
**Budapesti Műszaki és Gazdaságtudományi Egyetem**
Villamosmérnöki és Informatikai Kar
Távközlési és Médiainformatika Tanszék

# Vásárlói viselkedés előrejelzése hűségkártya-adatok alapján

TDK DOLGOZAT

*Készítette*
Kazi Sándor Antal
sandor.kazi@gmail.com

*Konzulens*
Gáspár-Papanek Csaba
gaspar@tmit.bme.hu

2011. október 28.

# Tartalomjegyzék

**Bevezető** **3**

**1. A feladat bemutatása** **4**

1.1. A feladat részletei . . . . . 4

1.2. Az adatok . . . . . 5

1.3. A feladat nehézsége, érdekessége . . . . . 10

1.4. Eszközök . . . . . 11

**2. Többcímkés tanulás (multi-label learning)** **12**

2.1. A feladat, mint multi-label tanulás . . . . . 13

2.2. Elért eredmények . . . . . 13

**3. Megközelítések a feladat megoldására** **15**

3.1. A csapat által leadott megoldás . . . . . 15

3.2. Saját megközelítések: alapvetés . . . . . 16

3.3. Statisztikai megközelítés . . . . . 16

3.4. Adatbányászati megközelítés . . . . . 18

3.5. Multi-label . . . . . 19

3.6. Vegyes megoldások . . . . . 20

**4. Eredmények és értékelés** **21**

4.1. Kiértékelés . . . . . 21

4.2. Statisztikai megoldás . . . . . 22

4.3. Adatbányászati megközelítés . . . . . 23

4.4. Multi-label megoldás . . . . . 24

4.5. Több megoldástípus ötvözése és összehasonlítás . . . . . 25

4.6. Fejlesztési lehetőségek és összegzés . . . . . 26

**Köszönetnyilvánítás** **28**

**Ábrák jegyzéke** **29**

**Táblázatok jegyzéke** **30**

**Irodalomjegyzék** **32**

## HALLGATÓI NYILATKOZAT

Alulírott *Kazi Sándor Antal*, a Budapesti Műszaki és Gazdaságtudományi Egyetem hallgatója kijelentem, hogy ezt a TDK dolgozatot meg nem engedett segítség nélkül, saját magam készítettem, csak a megadott forrásokat (szakirodalom, eszközök stb.) használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Hozzájárulok, hogy a jelen munkám alapadatait (szerző(k), cím, angol és magyar nyelvű tartalmi kivonat, készítés éve, konzulens(ek) neve) a BME VIK nyilvánosan hozzáférhető elektronikus formában, a munka teljes szövegét pedig az egyetem belső hálózatán keresztül (vagy autentikált felhasználók számára) közzétegye. Kijelentem, hogy a benyújtott munka és annak elektronikus verziója megegyezik.

Budapest, 2011. október 28.

_____

*Kazi Sándor Antal*
hallgató

# Bevezető

A kiskereskedelem egyre több területét hódítják meg az úgynevezett hűségkártya megoldások. Mivel ma már a legtöbb nagy láncnak van ilyen megoldása, ezért használatuk célja egyre kevésbé az ügyfél-lojalitás növelése, sokkal inkább az ügyfélviselkedés megismerése. A hűségkártyák segítségével már összekapcsolhatók az ügyfelek különböző időpontokban végzett cselekvései, például a vásárlás vagy a kedvezmények igénybevétele. Sokat jelenthet egy áruház vagy szolgáltató számára, ha nagy biztonsággal tudja előre, hogy ügyfelei mikor fognak ismét vásárolni vagy egy-egy szolgáltatást igénybe venni. Ezeket az információkat raktárkészlet, a készpénzgazdálkodás vagy munkatársak munkaidejének meghatározása során tudják felhasználni az üzleti intelligencia megoldásokat használó cégek.

TDK dolgozatom alapja egy, ebbe a témakörbe illeszkedő, de speciális feladat: a megoldandó probléma egy hűségkártya alapú tranzakciós adathalmazra épül. Egy áruház adatai (vásárlás dátuma és kifizetett pénzösszeg) alapján meg kell jósolnunk, hogy az egyes vásárlók legközelebb mikor fognak megjelenni és milyen összeget fognak a pénztárnál hagyni, azaz a feladat a két paraméter minél pontosabb becslése. A becslés hatékonyságát az fejezi ki, hogy az összes vásárló közül hány esetében tudjuk pontosan eltalálni a napot, illetve közelítőleg eltalálni a kifizetett pénzösszeget is. A probléma sajátossága, hogy a két megbecsülendő paraméter egyértelműen nem független, ennek kihasználása az előrejelzésben azonban korántsem triviális adatbányászati feladat.

Dolgozatom célja az ismertetett feladatra bemutatni több megközelítést, ezek sikerességét gazdag futási statisztikákon keresztül összehasonlítani. Célom továbbá ismertetni a több paramétert becslő (*multi-label*) felügyelt tanulási módszerek alapjait illetve azok párhuzamát a dolgozatban ismertetett feladattal. Kitérek az előrejelzési feladatot leghatékonyabban megoldó adatbányászati és statisztikai módszereim bemutatására, és ezek önálló alkalmazásához viszonyítom a több paramétert becslő eljárások eredményeit.

# Első fejezet

# A feladat bemutatása

Mint ahogy az a bevezetőben elhangzott, egyre fontosabbá válik, hogy információink minél kisebb részéből is minél több következtetést legyünk képesek levonni. Az ismertetésre váró feladat a vásárlói viselkedés legmeghatározóbb részeit emeli ki egy a végletekig egyszerűsített formában. Ezek a részek, hogy a vásárló mikor fog jönni és mennyit fog fizetni, nyilvánvalóan összefüggenek a korábbi vásárlásokkal, de ez a függés rendkívül összetett – hiszen az ember néha azért is újra elmegy vásárolni, mert elfelejtett megvenni valami fontosat. Az összefüggés nem triviális, ráadásul emberenként és időszakonként is változó. Általános esetben nagyon sok – gyakorlatilag végtelensok – féle adatot fel lehetne használni ezen legmeghatározóbb részek becslésére az egyszerű dolgoktól, mint a vásárló neme az egészen bonyolultakig. Egy bonyolultabb tényező lehetne: az aktuálisan akciós termékek közül melyek közül vásárol rendszeresen, azaz utóbbi egy évben havi átlagban elköltött $k$ pénzmennyiséget erre a termékre, vagy ezzel ekvivalensre, ha ebből aktuálisan készlethiány volt. Minden új tényező felvétele a következtetésbe tovább tudja javítani annak eredményességét, de ronthatja az általánosítóképességet, ráadásul egyre bonyolultabb modellt és ezzel együtt egyre hosszabb számítási időt eredményez.

A feladat rendkívül egyszerű adathalmazon van értelmezve, ezáltal alkalmassá válik arra, hogy már rövid idő alatt egy jó képet szerezhessünk az általánosíthatóságról. A tényezők egy ilyen kis részletének megfogása ezen kívül arra is jó lehetőséget nyújthat, hogy egy nagyobb, több információra építő megoldás részeként alkalmazzunk egy ezen a kis darabon született modellt. Több olyan megoldás ismert változatos területeken, amely kisebb, egyszerű elemekből áll össze, mégis felveszi a versenyt a legjobb monolitikus megoldásokkal is – ennek tipikus példái az adatbányászatban az ún. *Boosting* eljárások (AdaBoost, LogitBoost [1]), amelyet alkalmazva egyszerű építőkockák együttesével végezhetünk el akár meglepően bonyolult adatbányászati feladatokat is.

## 1.1. A feladat részletei

Az eredeti feladat egy, a brit Dunnhumby cég által meghirdetett nemzetközi verseny központi problémája volt [2]. A verseny 2011. szeptember 30-án ért véget, a hajrájában harmadmagam-

mal részt vettem, összeségében a 48. helyet sikerült megszereznünk a közel háromszáz csapat között. A hajrában való részvétel a feladat sajátosságaival, érdekességével megismertetett, segített meglátni az adathalmaz elemzési lehetőségeinek sokszínűségét. A verseny végeztével is van lehetőség becslések kiértékelésére, ezáltal az eredeti feladatnak megfelelő formában tudom a feladatot TDK dolgozatként kidolgozni. Dolgozatomban az egyes megközelítések esetében részletesen bemutatom az általam készített megoldásokat, emellett összehasonlítási célból kitérek a csapat által készített és beadott közelítések elvére és eredményeire is. Ismertetem, hogyan értem el a versenyen beadottnál jobb megoldást, amely még továbbfejleszthető.

A feladat egy tranzakciótörténeti adatbázison alapul. Az adatok egy évnyi vásárlást fednek le, ami 110000 vevőhöz köthető, az adathalmazban egy rekord pontosan egy vásárlásának felel meg. Egy rekord három attribútum együttese: a vevőt azonosító kód (*customer_id*), a vásárlás dátuma (*visit_date*) és az elköltött pénzösszeg (*visit_spend*) (bár nincs pénzösszeg megadva, a kiíró cég nemzetisége alapján valószínűleg fontról, dollárról vagy euróról beszélhetünk). A tranzakciótörténetben 2010. április 1-jétől 2011. március végéig szerepel a 110000 vevő összes vásárlási dátuma az összegekkel. A mintahalmaz egy igen nagy részének, 100000 főnek a további vásárlási adatait is ismerjük, egészen 2011. július végéig. A feladat a fennmaradó 10000 vevő következő vásárlása nézve minél pontosabb becslést adni. A becslés pontosságát az fejezi ki, hogy az esetek hány százalékában találtuk el a vásárlás napját úgy, hogy közben a fizetett összeg tekintetében is tízes intervallumon belül vagyunk. A dátumbecsléseket szintén a július végéig tartó intervallumból várja a rendszer és előírja, hogy a beadott tábla pontosan 10000 rekordot tartalmazzon – tehát nem lehetséges olyan predikciót beadni, amelyben azt feltételezzük, hogy a vásárló a megadott időszakban nem fog visszatérni.

A feladat „előrejelzés" mivoltának megfelelő logika alapján a predikció által érintett intervallumra a jövőbeli, a megelőző egy évre pedig a múltbeli vagy azzal egyenértékű megnevezést alkalmazom. Emellett a továbbiakban a pénzmennyiségre a dollár megnevezést is alkalmazom. A verseny kapcsán bevezetendő fogalom még a toplista. A verseny folyamán lehetőség volt naponta kétszer feltölteni a 10000 vásárlóra vonatkozó predikciót. A feltöltés alapján a 10000 egy kis szeletén (a fórum tanúsága szerint nagyságrendileg 1000 vásárló) végzett futtatás százalékos találati aránya alapján egy publikus toplistát határozott meg a rendszer. A verseny végeztével a kiértékelés már a 10000 rekord alapján történt, emiatt történt is egy helycsere a díjazott második és a nem díjazott negyedik hely között a verseny lezárultál. A toplista tesztelés véletlenszerű szeleten történt, emiatt ha a toplistára optimalizál a versenyző kisebb az esély az ún. túltanulás jelenségére.

## 1.2. Az adatok

Az adatfájlok méretével kapcsolatos információkat az 1.1. és az 1.2. táblázatok tartalmazzák. Előbbi az adatfájlok logikai méretével foglalkozik, míg utóbbi a hozzájuk tartozó intervallummal.

A feladat kiindulási alapja az 1.2. táblázatban felsorolt három fájlban tartalma. Látható, hogy mindkét adathalmazban átlagosan 100 körüli vásárlás tartozik tartozik az egyéves tartományhoz és több, mint 20 a kérdéses időszakhoz (csak a tanító adathalmazon ismert).

| | **tanító halmaz** | **teszt halmaz** | **összesen** |
|---|---|---|---|
| ügyfelek száma | 100000 | 10000 | 110000 |
| rekordok száma az első egy évben | 9950921 | 1008142 | 10959064 |
| rekordok száma a kérdéses intervallumban | 2195716 | – | – |
| rekordok száma összesen | 12146637 | – | – |

**1.1. táblázat.** *A rendelkezésre álló adatok mennyiségi jellemzői*

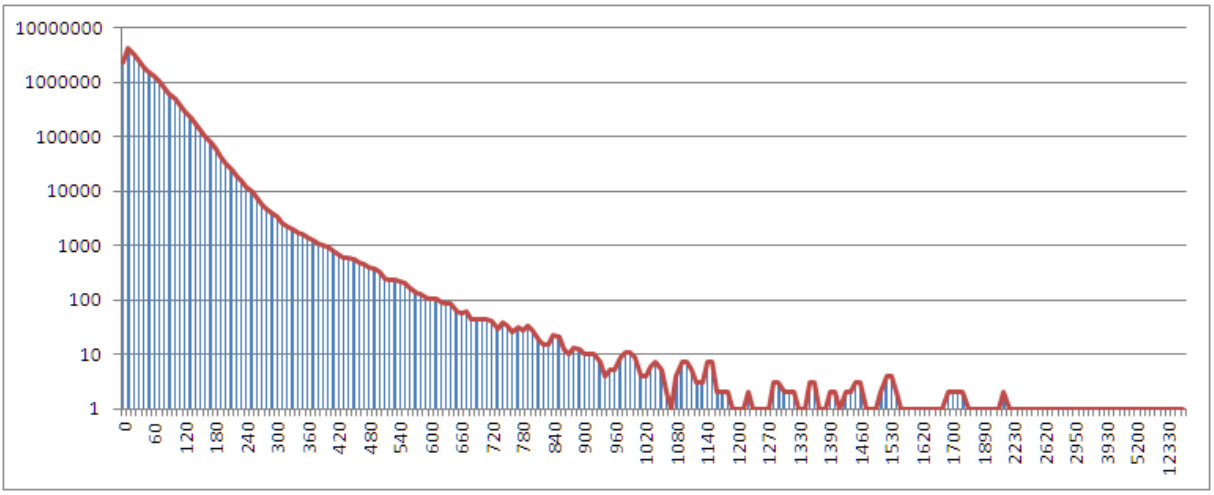
| **fájlnév** | **leírás** | **kezdő dátum** | **befejező dátum** |
|---|---|---|---|
| training.csv | tanító adathalmaz | 201**0**.04.01. | 2011.0**7**.31. |
| test.csv | kiértékelő adathalmaz | 201**0**.04.01. | 2011.0**3**.31. |
| example_entry.csv | predikció beadási minta | 201**1**.04.01. | 2011.0**7**.31. |

**1.2. táblázat.** *A rendelkezésre álló tranzakciótörténet részletek időhatárai*

A vásárlások szezonlitását mutatják be az 1.1.-1.9. ábrák. Több grafikon esetében is igaz, hogy a tengelyek értékkészlete nem folytonos, a grafikonok mégis folytonos vonallal rajzoltak. Ennek oka, hogy csak a diszkrét ponthalmazokat szerepeltetve a grafikonokon azok nem szemléltették ennyire látványosan a hasonlóságokat és különbségeket.

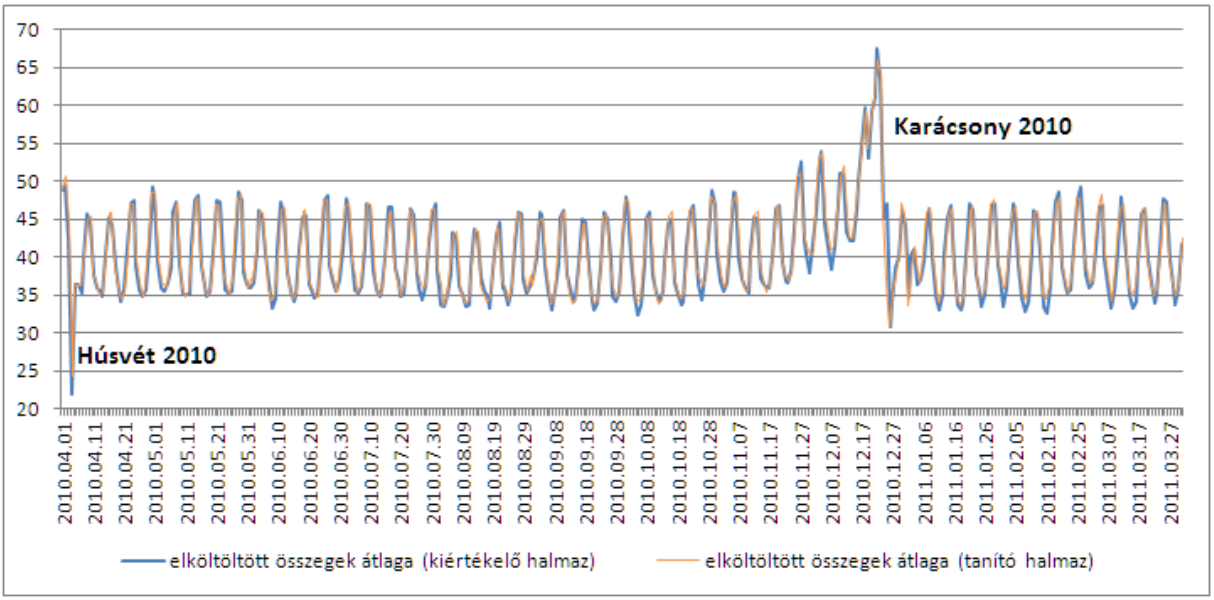
Az 1.1. ábrán szereplő grafikon a kifizetéseket ábrázolja csoportosítva. Tízdolláronként felfelé haladva szerepel minden értékhez a vízszintes tengely mentén, hogy a megelőző egy évről rendelkezésre álló adataink alapján hány olyan vásárlás történt, ahol a kifizetett összeg az adott érték tízdolláros környezetébe esik. A lépésméret miatt minden egyes vásárlás legfeljebb két intervallum-középpont esetében szerepelhet. A vásárlások száma logaritmikus skálán van ábrázolva. Mivel a függvény alakja ezen a grafikonon logaritmikus jellegű, ezért az eredeti függvény log log jellegű. Az ábrán látható, hogy az alacsony összegek relatív gyakorisága viszonylag következetes, és minél magasabb összegről beszélünk, annál inkább véletlenszerű.

Az 1.2. és az 1.3. ábrákon az egyes napokhoz tartozó vásárlások átlagértéke illetve mediánja látható. Az ábrákon megfigyelhető, hogy a hetek közel periodikussá teszik ezt a grafikont. Észrevehető ezen túl, hogy a vásárlási értékek a 2010-es húsvét és karácsony esetében jelentősen eltérnek a szokványos értékektől. Mindkét grafikonon a megelőző egy év vásárlásai láthatók a tanító és a kiértékelő halmaz adatait figyelembevéve. A grafikonokon az is jól látszik, hogy kevesebb rekordot tartalmazó kiértékelő adathalmaz hátterében is (nagy valószínűséggel) hasonló eloszlás húzódik meg, azonban a minta mérete miatt az eloszláson látható zaj nagyobb. Amennyiben háttérben tényleg ugyanaz az eloszlás rejtőzik, akkor a kiértékelő minta esetében hozott becsléseinknél felhasználhatjuk a tanítóminta bizonyos részeit is a következtetéshez. Ez a két mérték azért lényeges, mert a két leggyakrabban használt fogalom egy mérték általánosítására. A predik-




**1.1. ábra.** *a megelőző egy év vásárlásaikor kifezetett pénzösszegek közül ennyi esik egy-egy érték tízdolláros környezetébe (a tanító és a kiértékelő adathalmazon együttesen) – a grafikon értelmezési tartománya az ábrán láthatónál bővebb, a nagyobb értékek szemléltetési célból hiányoznak*

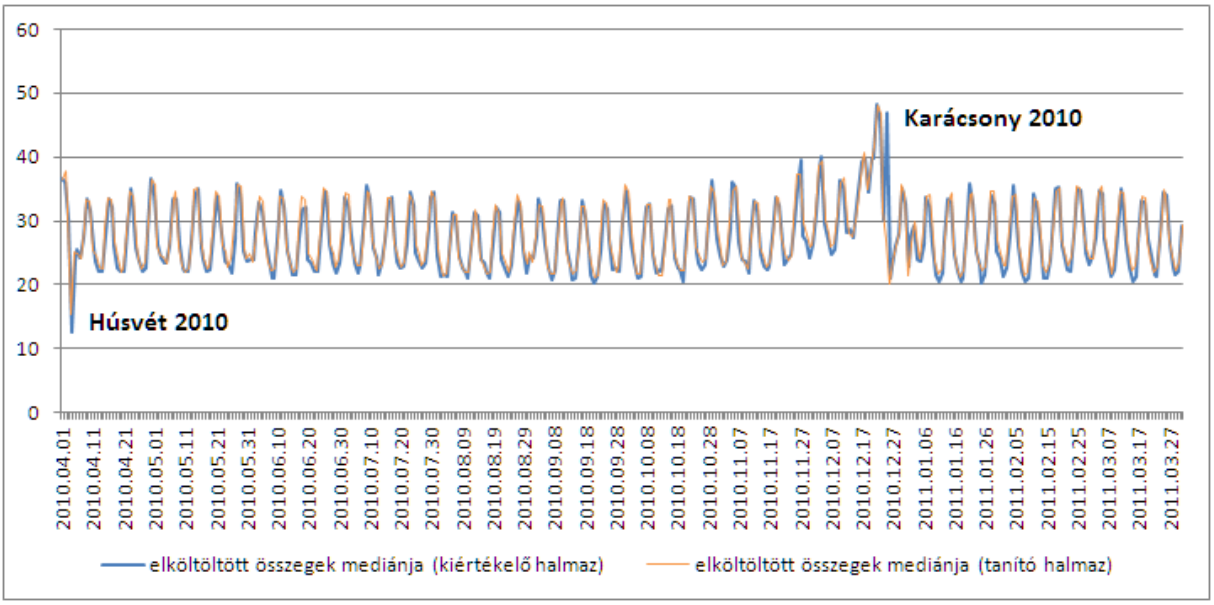
ció esetében jól használható lehet mind az átlag (számtani közép) mind pedig a medián (sorrend szerinti középső érték).



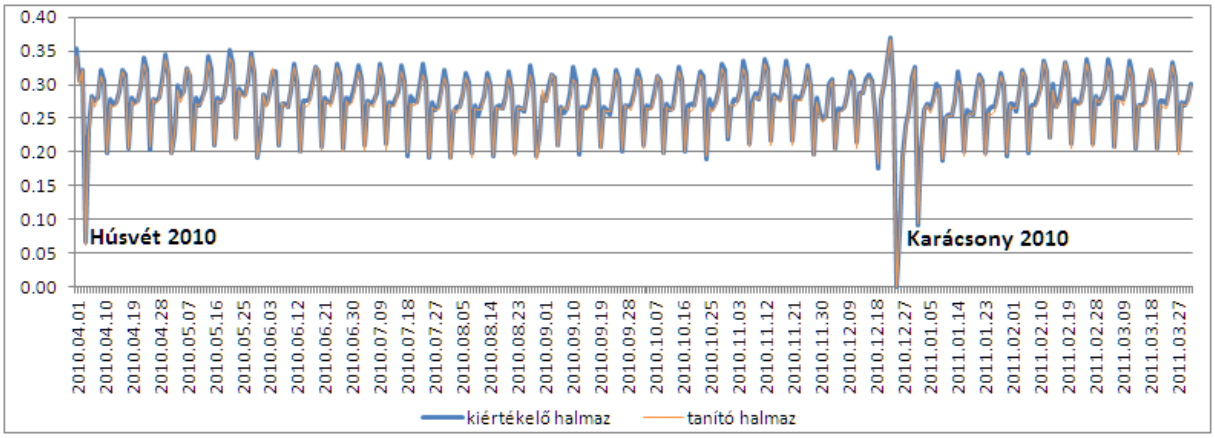

**1.2. ábra.** *a megelőző egy év egyes napjaihoz tartozó elköltött összegek átlagértéke (összehasonlítva a tanító és a kiértékelő adathalmazon)*

Az 1.4. és az 1.5. grafikonokon az egyes napokon betérő vásárlók számának alakulása figyelhető meg a megelőző egy évről illetve kérdéses nyolcvannapos intervallumról. Előbbi a kétféle adathalmaz együttese, míg a második ábra csak a tanító halmaz (csak az áll rendelkezésre) alapján készült. Mindkét grafikonom megfigyelhetők a hetes ismétlődések, ahogyan ez látható volt az átlagosan kifizetett értékek esetében is.

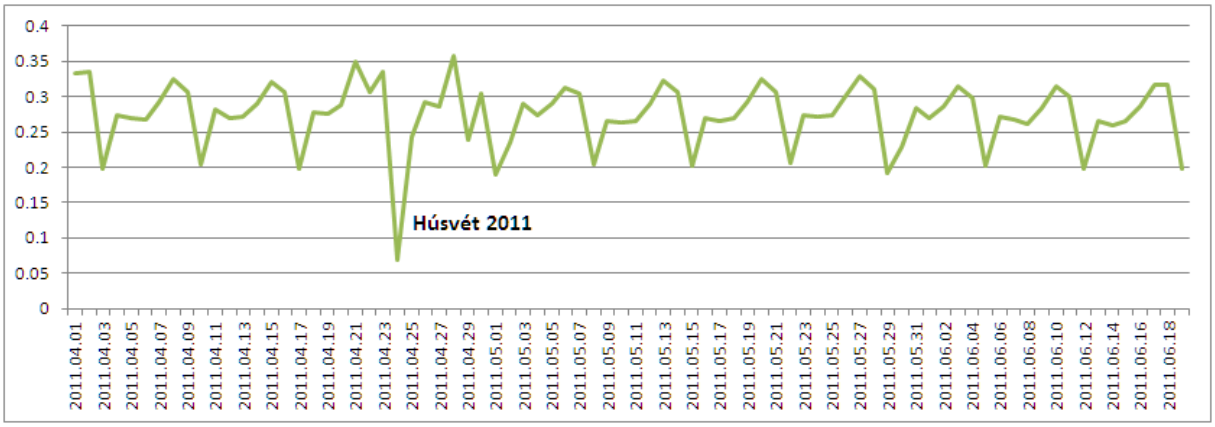
Az 1.6. és az 1.7. ábrákon szereplő grafikonokokon szintén megfigyelhető a hetes közelítő periodicitás és a két grafikon közel azonos mivolta. A grafikonokon az egyes napokhoz tartozó




**1.3. ábra.** *a megelőző egy év egyes napjaihoz tartozó elköltött összegek mediánja (összehasonlítva a tanító és a kiértékelő adathalmazon)*

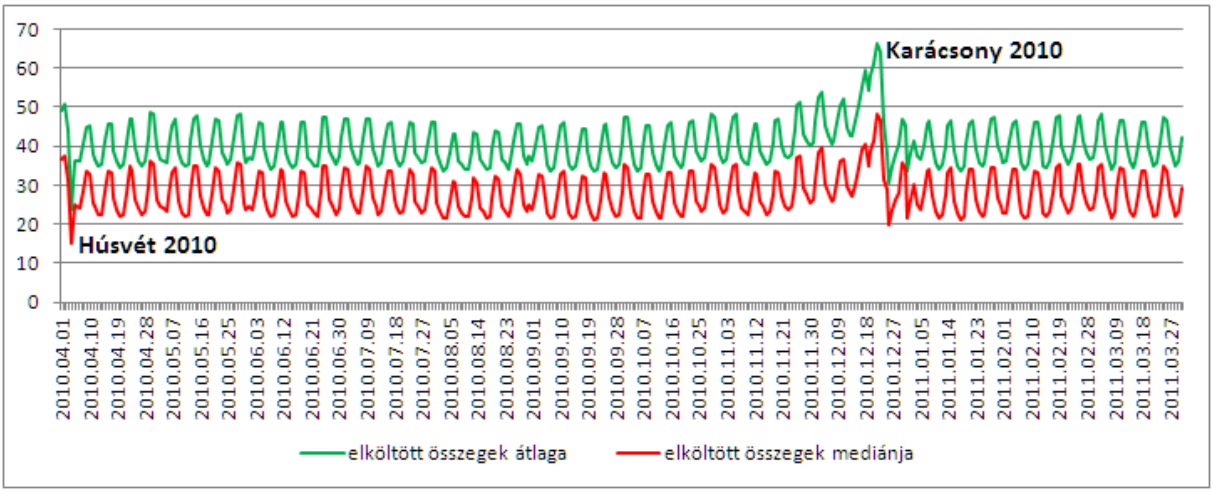



**1.4. ábra.** *a megelőző egy év egyes napjaihoz tartozó vásárlások száma (a tanító és a kiértékelő adathalmazon együttesen)*

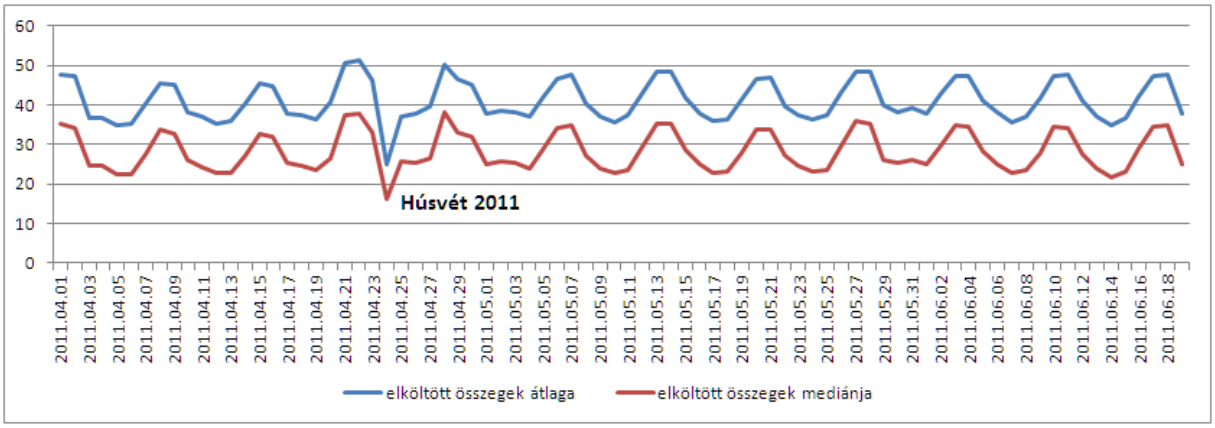



**1.5. ábra.** *a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó vásárlások száma (a tanító adathalmaz adatai alapján)*

vásárlások átlagának és mediánjának összehasonlítása látható a megelőző egy éves és a soron következő nyolcvan napos időintervallumra vetítve. Látható, hogy az ünnepnapok az ábrákon jelölt helyeken a szokványoshoz képest magasabb, majd nagyon alacsony értékeket eredményeznek.

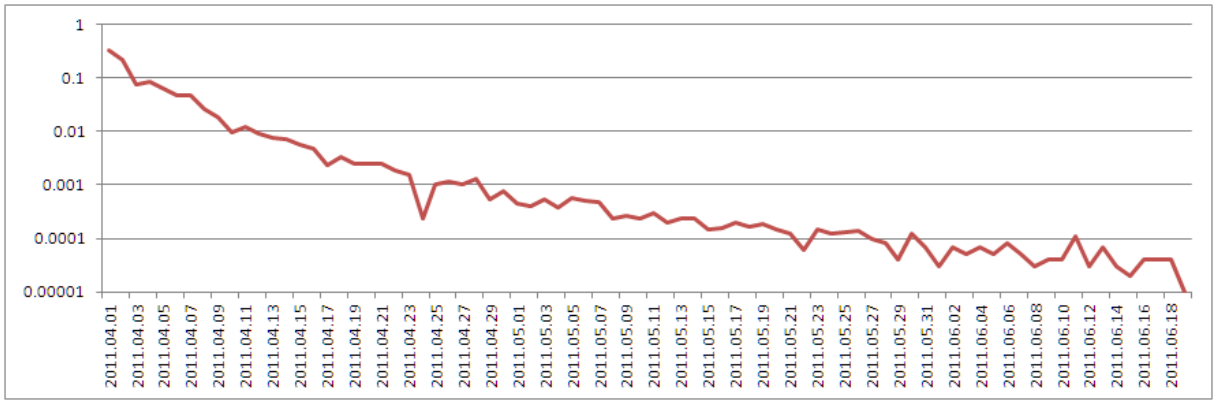



**1.6. ábra.** *a megelőző egy év egyes napjaihoz tartozó vásárlások átlagértéke és mediánja (a tanító és a kiértékelő adathalmazon együttesen)*

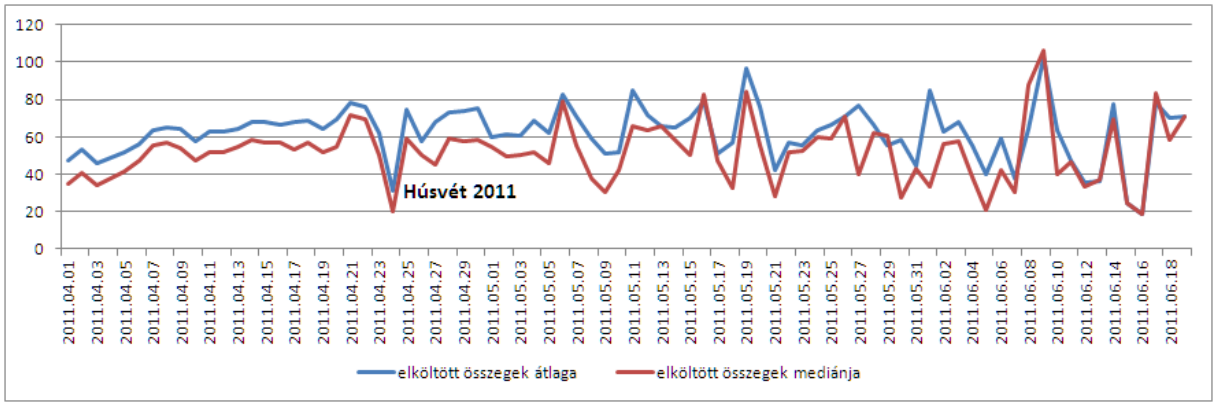



**1.7. ábra.** *a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó vásárlások átlagértéke és mediánja (a tanító adathalmaz adatai alapján)*

A tanítóhalmazból leválogatva a következő vásárlásokat – azaz az értékeket, amelyekre a predikciót kell szolgáltatni – több dolog is feltűnhet nekünk. Ezek megértését segíti az 1.8. és az 1.9. ábra. Az ábrákon az látható, hogy a soron következő első vásárlások átlagosan és medián tekintetében milyen összegben történnek, valamint, hogy az egyes napokon hány vásárló teszi meg első vásárlását 2011. március 31. után. Látható az 1.9. ábrán, hogy az idő előrehaladtával egyre nagyobb intervallumból kerülnek ki az átlag és a medián értékek. Az átlagos és medián pénzösszeg grafikonon látható, hogy az idő előrehaladtával elinte egyre magasabb az átlagérték. Még kivehető a húsvét környéki törés, majd, ahogy a vásárlók száma már nagyon lecsökken kevésbé egyenlítődik ki, látszólag „véletlenszerűvé" válik az átlagérték. Ebből arra is következtethetünk, hogy valószínűleg a korábbi vásárlásokat többségében könnyebb előrejelezni, mint a későbbieket.




**1.8. ábra.** *a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó első vásárlások száma (a tanító adathalmaz adatai alapján)*




**1.9. ábra.** *a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó első vásárlásokhoz tartozó pénzösszegek átlaga és mediánja (a tanító adathalmaz adatai alapján)*

Érdekesség, hogy a tanítómintában szereplő 100000 vásárló közül csak 99952 esetében láthatunk az adathalmazban következő vásárlást, 48 nem tért vissza a megadott intervallumban a bevásárlóközpontba. Amennyiben a tanító és a teszt halmaz esetében megegyező eloszlást feltételezünk, akkor a 10000 vásárló közül hozzávetőlegesen 5 vásárlót hiába várunk – tehát, mivel olyan predikciót nincs lehetőség adni, hogy a vásárló nem tér vissza, hibátlan predikció valószínűleg nem is adható be.

## 1.3. A feladat nehézsége, érdekessége

A feladat az egyszerű megfogalmazhatósága ellenére igen bonyolult összefüggések feltárását célozza meg. A vásárlók viselkedését nap mint nap annyi tényező befolyásolja, hogy ha vannak is jól formalizálható vásárlási szokásaik, az adathalmazban azok tiszta ábrázolásához képest nagymennyiségű zajt is tapasztalhatunk. Ez a torzító tényező jelen van mind a bemeneti, mind a kérdéses adatokon: a cél a zajos minták alapján a zajos eredmény becslése. Azonban mivel ez a torzító erő időben és emberenként is változik, ezért még nehezebb feladat a tényleges értéket megbecsülni. Az online tesztelés hátulütője, hogy amíg a verseny szervezői nem teszik lehetővé,

a kiértékelő halmazon a kategóriánkénti illeszkedés eredménye is ismeretlen. Várható azonban, hogy ez a változtatás megtörténik.

A feladat attól válik igazán érdekessé a maga nemében, hogy két értéket kell megbecsülni párhuzamosan, amelyek ráadásul nem függetlenek. Lehetséges, hogy a két paraméter külön-külön becsült értéke alacsonyabb egy összességében jobb eredményt elérő algoritmus esetében. A két érték összefüggése meglehetősen bonyolult, de már egy egyszerű példán át is szemmel látható: hiszen ha egy nap elteltével ismét ellátogat egy vásárló a bevásárlóközpontba, akkor igen nagy valószínűséggel kevesebb dolgot vesz, emiatt valószínűleg a kifizetett összeg is alacsonyabb lesz, mint ha öt nap kihagyással vásárolt volna újra. Tehát az összefüggést nem csupán ezt a két érték határozza meg, hanem erősen függ a vásárlótól, az időponttól illetve sok egyébtől, amelyek feltárása már a feladat megoldásának részét képezi.

## 1.4. Eszközök

A feladat megoldására és a dokumentáció elkészítésére több informatikai eszközt is felhasználtam. Az adatok kezelésére alapvetően az SQL nyelvet alkalmaztam egy PostgreSQL 8.4 adatbázisszerver segítségével. Nagy segítséget jelentett az ún. „ablakfüggvények" (*window functions*) felhasználhatósága – más környezetben is létezik hasonló, Oracle esetében analitikus függvényeknek (*analytic functions*) nevezik a nagyon hasonló funkcionalitású speciális függvényeket. Az adatok betöltésére, közvetetten az algoritmusok futtatására és az exportálásra bash shell scripteket készítettem. Ezeken kívül az ingyenes RapidMiner [3] alkalmazás legfrissebb változatával ismerkedtem meg és használtam fel arra, hogy az adatbányászati algoritmusokat teszteljem és futtassam. A RapidMiner szoftver megismerését sok online videó *tutorial* segíti, ezek nagy segítségemre voltak.

A dokumentációt a LaTeX dokumentumszerkesztő nyelven készítettem el a TeXnicCenter [4] és a MikTeX [5] szoftverek támogatásával. A teszteléshez a rendelkezésre álló adatokat és a verseny hivatalos honlapján [2] elérhető, még a verseny lezárása után is szabadon használható kiértékelő rendszert használtam.

# Második fejezet

# Többcímkés tanulás (multi-label learning)

A *multi-label learning*, amelyet többcímkés tanulás névre fordíthatunk („többcímkés" elnevezés: [6]), azt a problémakört takarja, amikor egy-egy entitás kapcsán több kimeneti dimenzió is megadásra vár. Egy egyszerű, a többcímkés tanulás témakörében gyakran alkalmazott példa, ha filmeket szeretnénk kategóriákba sorolni. Ilyenkor mindegy egyes filmre igaz az, hogy a kategóriák közül többe is beletartozhat – azaz egy-egy kategóriát bináris osztályozási feladatként értelmezve nem diszjunk osztályokról beszélünk. Tehát nem arról van szó, hogy a különböző címkék az egyes osztályok, és azokba sorolunk, sokkal inkább arról, hogy a lehetőségek között a hatványhalmaz elemei szerepelnek, a címkekombinációk. Talán az egyik legkézenfekfőbb példákkal élve: lehet egyszerre egy film sci-fi és akciófilm is, megkaphatja mindkét címkét, ugyanígy kezelendő a romantikus vígjáték kategória is.

Ugyanezen a példán keresztül könnyedén bemutatható, a *multi-label* algoritmusok egy kitüntetett halmaza, az átfogó metódusok – szó szerinti fordításban „együttes" metódusok (*ensemble methods*). Tegyük komplexebbé a filmcímkézés modelljét a priori feltételezésekkel a címkék közötti összefüggések tekintében. Feltehetjük például, hogy csak nagyon ritkán lehet szükség arra, hogy egy filmről megállapítsuk, hogy az a romantikus-akciófilm, romantikus-horror, dráma-vígjáték, krimi-vígjáték vagy thriller-mese kategóriapárokba egyszerre tartozik, viszont ezek nagyrészét nem zárhatjuk ki teljes biztonsággal. Átfogó metódusnak (*ensemble method*) nevezhetjük azokat a módszereket a többcímkés tanulás témakörén belül, melyek felhasználják az egyes címkék közötti összefüggéseket, azaz egyszerre több címke értékét veszik figyelembe. Azokat a problémákat pedig, amelyek az ilyen típusú metódusok alkalmazására terepet biztosítanak, ezzel párhuzamosan elnevezhetjük átfogó többcímkés tanulási problémáknak.

Már a bemutatott egyszerű példa is remekül mutatja, hogy sok esetben logikusan alkalmazható, a modell részeként értelmezhető elem a címkék közötti összefüggés. Igazán nagy helyzeti előnyt természetesen akkor jelenthet *multi-label* módszerek alkalmazása, ha a teljes címkézési sikerességet (például egy filmen az összes címke stimmeljen) szeretnénk maximalizálni, nem pedig

egymástól függetlenül a címkénkénti sikeres osztályozást. Ezen belül is leginkább akkor, ha a címkéken külön-külön jól teljesítő algoritmusok összeségében látszólag rosszul teljesítenek. Ilyenkor a címkék közötti korreláció kiaknázásával lehetőségünk lehet a összesített sikerességet növelni.

## 2.1. A feladat, mint multi-label tanulás

Jelen feladat esetében is beszélhetünk korrelációról a két címke között, mint ahogy arról már a feladat bevezetésekor, a 1.3 részben említést tettem. Az összefüggések figyelembe vételére szükség lehet, hiszen a két érték egymástól független becslése nem feltétlenül vezet eredményre. A *multi-label* megközelítés ezen felül logikusan következik is a feladat elvárásaiból. Hiszen a maximalizálandó érték a két címke helyes osztályozási metszetének aránya a teljes mintában. A két címke közötti összefüggés ebben a feladatban az idő-pénz irányban egyszerűen megfogható: próbáljuk meg a vásárlási időre adott predikciót is felhasználni a pénzösszeg megbecslésénél. Ezt semmi nem tiltja, ezen felül a modellünkhöz csak plusz információt adhat a célfüggvény tekintetében, hiszen csak azon rekordok esetében számít a pénzösszeg tippjeink pontossága, amelyeknél a dátumot eltaláltuk. Tehát ha az összegbecslésnél a predikált dátumot belekalkuláljuk a modellbe, azzal hamis információkat nem viszünk bele.

A másik irány, azaz a dátum függése a predikált összegtől már nem is ennyire természetes, illetve pontatlan információkat vihet a modellbe. Nem annyira természetes, hiszen, hogy mennyit fizetünk, leggyakrabban akkor dől el, amikor már az áruházban vagyunk. Ráadásul a szupermarket tevékenységének sokszor az a célja, hogy minél több vásárló vegyen meg minél több olyan terméket, amelyet előre nem tervezett be, ezáltal pedig a pénzösszeg is megváltozik. Emellett ez az irány pontatlan információkat is visz a modellbe, hiszen ha mondjuk 20 dollár a tippünk, az akkor is találat, ha 10 dollárért vásárol a vevő és akkor is, ha 30-ért. Egy háromszoros különbség már meglehetősen különböző következtetéshez kellene vezessen, de ez a modellből nem fog kiderülni.

Mindezek alapján valószínűleg érdemesebb a már kész dátumbecslést felhasználni a pénzösszeg predikciójánál, a fordított irányú következtetéseket pedig kihagyni a modellből.

A feladat esetében nem teljesen egyértelmű, hogy regresszióról vagy osztályozásról beszélhetünk. Hiszen a dátum becslése osztályozási feladat – nem számít mennyivel tévedünk, pontos érték kell – viszont a pénzösszeg becslésénél már számít a tévedési mérték, minél közelebb szeretnénk kerülni. De a pénzösszeg becslése sem egyértelmű regresszió, hiszen csak a tíz dollár sugarú intervallumba kerülés számít, ez pedig már megfogalmazható osztályozási feladatként is.

## 2.2. Elért eredmények

*Multi-label* tématerületen több irodalom is szövegbányászati feladatokkal foglalkozik, azok kapcsán merül fel a többcímkés megoldások alkalmazása [6] [7] [8]. A példaként felhasznált filmcímkézési feladat is tartozhat ebbe a kategóriába: lehetséges például, hogy az egyes filmek leírásai

alapján akarjuk azokat kategóriákba sorolni szövegbányászati módszereket használva. Sokféle tématerülettel kapcsolatban kerül elő ez a fogalom, sokféle feladatra nézve. Ilyen például a képek címkézése [9], amely egy képfeldolgozási problémát, illetve szövegbányászati alkalmazást vet fel, amennyiben a képhez egy kapcsolódó szövegkörnyezet vagy szöveges leírás is rendelkezésre áll. Létezik kísérlet zenék hangulatokhoz rendelésére automatizáltan, amely szintén többcímkés problémáként vizsgálja saját feladatát [10]. De vannak orvosi, genetikai kutatási témák is, amelyek a többcímkés osztályozás problémájával foglalkoznak [11] [12]. Látható tehát, hogy a *multi-label* osztályozás és regresszió problémájával több területen is szembekerülhetünk. Szintén látható, hogy a problémák ily módon történő megközelítése sok esetben kifizetődő volt.

Találhatók a témában tématerülettől független, elméleti kutatási eredmények is. Ilyen eredmény az LP, azaz a címkehatványhalmaz (*label powerset*) megközelítés, illetve a *random k-labelsets* [13]. A témakörfüggetlen eredmények kis aránya az összes kutatási eredmény között azt is mutatja, hogy sok esetben hasznos a feladattípus, témakör illetve maga a feladat szempontjából specifikus *multi-label* megoldások elkészítése. A sikerekre jó példa egy korábbi verseny eredménye is: a KDD Cup 2009-es versenyén győztes IBM Research csapata is *ensemble* metódust használt a győztes módszerében [14].

# Harmadik fejezet

# Megközelítések a feladat megoldására

Predikcióról lévén szó, nem egyértelmű, hogy a feladat megoldásához milyen módszerrel kell hozzákezdeni, hogy a legjobb eredményt érjük el. Még az sem teljesen evidens, hogy a problémát milyen témakörbe illeszkedően próbáljuk meg megfogalmazni. A feladatra egyaránt mondhatjuk, hogy az statisztikai, adatbányászati, gépi tanulási vagy akár azt, hogy pszichológiai, viselkedéselemzési probléma. Emiatt a rendelkezésre álló eszközökből való választás már csak azért sem triviális, mert azok listája sem jól meghatározott. Az általam elkészített algoritmusok egy része statisztikai, egy másik része a gépi tanulás eszközeit veti be adatbányászati módszerekkel karöltve. Mindenezek segítségével elkészítettem *multi-label* megoldásokat is, valamint a megoldásaim ötvözésével még további predikciókat. TDK dolgozatomban ezek közül a legsikeresebbeket ismertetem, és ezeket össze is hasonlítom a verseny lezárása előtt beadott, még csapatban készített megoldással.

## 3.1. A csapat által leadott megoldás

Röviden ismertetem, hogy a verseny éles szakaszában a csapatom mi alapján adta be az akkori végleges megoldást. Az én megoldásaim nem ebből indulnak ki, de jó összehasonlítási alapot biztosítanak a versenyen elérhető helyezés és a többi mérőszám tekintetében is.

A beadott megoldásban a két érték becslése egymástól függetlenül történt. A becsült dátumértékek a $k$-NN ($k$-közeli szomszéd – *$k$-nearest neighbor* [15]) módszer segítségével kaptuk, úgy, hogy gyakorlatilag az első hét napra adtuk az összes tippünket. A $k$-NN algoritmus a hetek hasonlósága alapján találta meg azt, hogy melyik hét lesz várhatóan a legjobb közelítése a soron következőnek. Mindezt vásárlónként külön-külön becsültük, tehát a háttérben 110000 modellt alkottunk meg. Mivel 2011. április 1-je péntek volt, ezért a $k$-NN szempontjából a hetet a péntek-csütörtök intervallummal adtuk meg. A hét hét napján kívül volt egy nyolcadik lehetőség is, mint tipp, ha ezt kaptuk, akkor egyszerű szabályalapú módszerrel megtippeltük a következő érkezést az egyhetes ablakon kívül, a korábban két vásárlás között eltelt idő legjellemzőbb értékei alapján.

A pénzösszeg becslésénél egy valamivel egyszerűbb megoldást választottunk. Megnéztük,

hogy az adott vásárló utolsó vásárlása milyen napra esik, majd az összes korábbi, ilyen napot követő vásárlások összegeinek összessége közül a mediánt választottuk ki.

## 3.2. Saját megközelítések: alapvetés

Több megállapítást tehetünk az adathalmazra és a predikciós elvekre nézve, ami alappillére lehet egy jó megoldásnak. Ezek a megállapítások azt a célt szolgálják, hogy az adathalmaz azon részeit óvatosan kezeljük, amelyek a kérdéses időszakra nézve nehezen használhatók, illetve ha egyértelműen csak javíthatunk bizonyos döntések mentén az elérhető eredményen, akkor ezeket a döntéseket hozzuk meg. Megvizsgálva a feladatot és az adathalmazt a következő megállapításokat tehetjük:

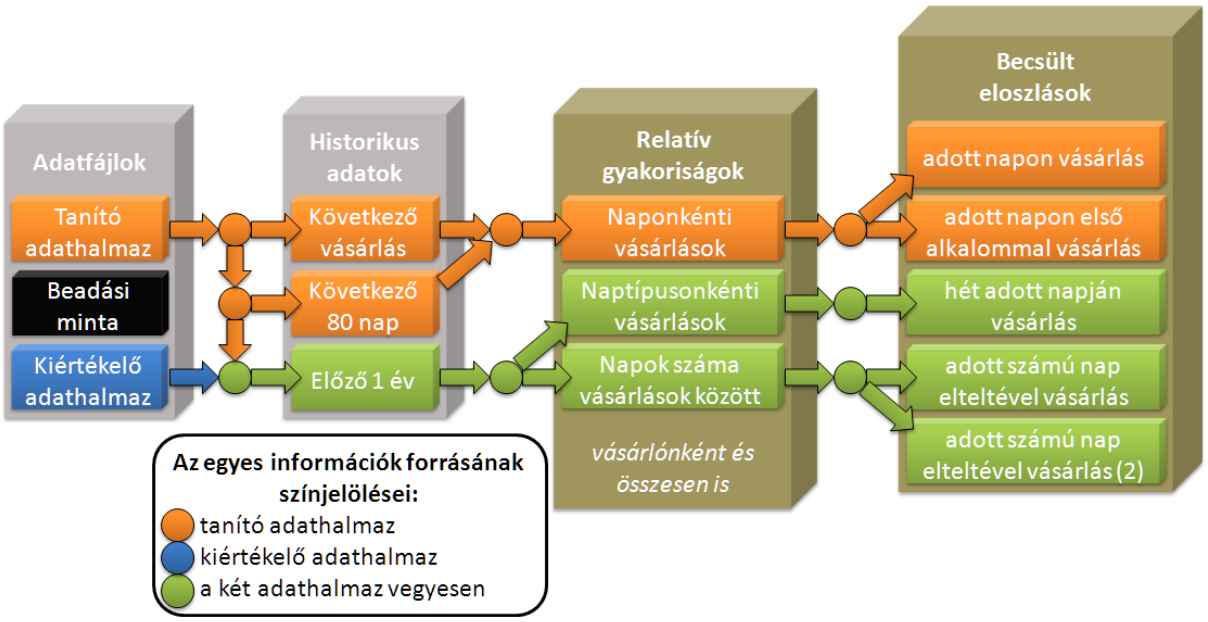
- az adathalmazon felismerhetők az egyhetes időszakok, tehát az egyes hetek, illetve azonos nevű napok között bizonyos szinten párhuzam vonható
- az egyhetes időszakok hasonlóak egy éves távlatban is
- a hét napjai között különbségek láthatók
- a karácsonyi időszak láthatóan jelentősen különbözik a többitől
- a húsvéti időszak láthatóan jelentősen különbözik a többitől (mindkét évben)
- 10 dollár alatti tippet nem érdemes tenni, mivel negatív összegek nincsenek – emiatt ha 10-re tippelünk a 10 alatti tippek helyett, csak nyerhetünk találatokat, elveszteni egyet sem tudunk

## 3.3. Statisztikai megközelítés

Megközelíthető a probléma pusztán statisztikai alapokon. Elsőként egy ilyen megközelítést mutatok be a probléma megoldására. Ebben az esetben szükség van további megállapításokra, amelyek mentén következtetünk. Egy tisztán statisztikai megközelítés alapja lehet, ha feltételezzük hogy a tanítóminta hátterében húzódó eloszlás megegyezik a tesztmintáével azon a szakaszon is, amely számunkra ismeretlen, vagyis amelyen a becsléseket kell szolgáltatnunk. Ezt az 1.2. rész alapján nyugodtan megtehetjük, látványosan egybeestek az értékek.

Ahhoz, hogy arról beszélhessünk, hogy a legközelebbi vásárlás bekövetkezése melyik napon a legvalószínűbb, valamilyen módon becsülnünk kell annak eloszlását. Ez a becslés többféleképp történhet, alkalmazhatunk ismert matematikai formulákat a probléma megoldására, vagy saját feladatspecifikus megoldást. Sűrűségfüggvény becslésére egy ilyen közismert módszer a magfüggvényes becslés (más néven: Parzen-Rosenblatt ablak [16] [17]), amely valamely kitüntetett sűrűségfüggvény paraméterezésével igyekszik a tényleges sűrűségfüggvény egy közelítő képét elérni, ezt alkalmazta dátumbecslésére a későbbi győztes, D'yakonov Alexander [18], aki ugyan megnyerte a versenyt összesítést, de éppen a dátumbecslésben nem jutott igazán magasra a toplistán.

Sok ehhez logikailag hasonló megközelítés létezik, például ha eloszlásokat készítünk a rendszerben fogalmilag létező, a modellben jó eséllyel fontos szerepet játszó tényezőkről, majd ezen eloszlások kummuláltja alapján adjuk meg a becslést az értékekre. Statisztikai megoldásaimnál ez utóbbira támaszkodtam.




**3.1. ábra.** *az általam készített statisztikai megközelítés dátumbecslésének döntéshozatalához szükséges információk előállítása – a relatív gyakoriságokat vásárlónként és összességében is előállítottam*

A vásárlási napjának predikciójában legjobbnak bizonyuló algoritmuscsoport a 3.1. ábrán követhető módon állít valószínűségbecsléseket, majd ezek alapján választja ki a legmegfelelőbb tippeket. Első lépése, hogy a jelenlegi fájlstuktúrát három halmazra bontja fel: a tanító adathalmaz és a kiértékelő adathalmaz rekordjai vegyesen az előző egy évről, a következő nyolcvan napra eső vásárlások a tanító adathalmazból illetve az utóbbi adathalmaz azon rekordjai, amelyek valamelyik vásárló első vásárlásához tartoznak 2011. április elsejétől kezdve. Az ábrán látható, hogy a második két csoport a tanító halmazból áll elő (hiszen csak abban vannak abból az időszakból információk), viszont az első csoport a két halmaz elegyítéséből. A három csoportból relatív gyakoriság kimutatásokat készítettem.

A jövőre vonatkozó két historikus adatcsoportból a naponkénti vásárlások és a naponkénti első vásárlások relatív gyakoriságát számítottam ki. Ez azt jelenti, hogy a nyolcvan napos kérdéses intervallumra megállapítottam minden naphoz egy-egy értéket, amelyek azt fejezik ki, hogy az adott nap milyen gyakorisággal volt a tanítóhalmazban valaki vásárlásának illetve első vásárlásának napja. Egy nagyon hasonló kivonata készül a múltbéli adatoknak is. Kigyűjtésre kerül egyrészt, hogy egyes naptípusonként (a hét napjai) milyen relatív gyakorisággal vásárol egy-egy vásárló az adott napon. Másrészt, hogy milyen relatív gyakorisággal fordul elő, hogy adott számú nap kihagyás után tér vissza a bevásárlóközpontba. Ugyanígy előállítottam egy olyan eloszlást is, amely a korábbi lépésméret függvényében adja meg vásárlónként, hogy a következő kihagyás

hány napig tart (az ábrán ezt jelöltem a (2) indexszel). Ezek alapján a relatív gyakoriságok alapján választottam ki a tényleges predikciót. Többféle módon is megpróbáltam összevonni a relatív gyakoriságokat.

A pénzösszeg becslésénél is eloszlást készítettem. Felhasználtam az 1.1. ábra elkészítésénél alkalmazott 10-dolláronkénti felbontást. A pénzösszeget a módusz alapján választottam ki, azaz amelyik a leggyakoribb volt, az lett a predikció.

Ezután mindkét becsült érték esetében súlyoztam az információkat és a relatív gyakoriságból így egy „hozzávetőleges valószínűséget" alkottam meg, amely a már súlyozott relatív gyakoriságból következik. A súlyozási és az időbecslés szerinti különböző összegzési metódusok által egy algoritmus többféle változatáról beszélhetünk, vagyis egy algoritmuscsoportról. Valamelyest javulást sikerült elérni azzal is, hogy a nagyon zajos részeket a mintából kihagytam, azaz a karácsonyi környéki és húsvét környéki rekordokat az eloszlásbecslésekbe nem számoltam bele. Ebben az esetben ügyeltem arra, hogy a hetes periodicitás megmaradjon, csak egész heteket hagytam figyelmen kívül. Az egyes algoritmusok súlyozási és összegzési módszereit és az az által elért eredményeket a következő fejezetben ismertetem.

## 3.4. Adatbányászati megközelítés

A csapat által beadott becslés a dátumra $k$-NN-t használt, emiatt már adatbányászati megoldásnak tekinthető. Már a csapat megoldásának elkészítésekor is kiderült, hogy az adathalmaz, ebben a historikus formában még nem igazán alkalmas adatbányászati algoritmusok alkalmazására. Egy lehetséges megközelítés, ha minden felhasználához hozzárendelünk egy-egy rekordot, amelyben az ő viselkedését leíró fontosabb információk találhatók, azaz aggregátumokat készíthetünk a historikus adatok alapján. Ezáltal egy felhasználónkénti adathalmazhoz jutunk, amelyen már futtathatunk gépi tanulási módszereket is. Az alapötlet, hogy az aggregátumokba kerüljön bele valahány korábbi vásárlás napja a legrégebbitől számozódva.

Több aggregátumot is előállítottam, mint:

- minimális vásárlási érték
- maximális vásárlási érték
- átlagos vásárlási érték
- vásárlási értékek mediánja
- átlagtól való négyzetes eltérés átlaga
- mediántól való négyzetes eltérés átlaga
- utolsó hat vásárlás napja (a hetedik legutolsó vásárlástól visszafelé számolva)
- utolsó hét vásárlás pénzösszege

- minimum nap a következő vásárláshoz (2011. április 1-ből következik) – a tanítómintáknál szükség szerint eloszlás alapján generált érték

A feladatot regresszióként értelmeztem: az aggregátum alapján tanítottam egy-egy SVM-et (szupportvektor gép, *Support Vector Machine* [19] [20], majd az elkészített modell alapján pedig megállapítottam a nap és a pénzösszeg predikcióját. A RapidMiner alkalmazásban futtattam egy ennek megfelelő processz elrendezést. RapidMiner kontextusban processznek nevezzük a processzelemekből (adatfeldolgozó, betöltő, modellépítő, modell alkalmazó, stb.) és vezetékekből felépített hálózatot. A processz a korábban SQL segítségével összeállított és exportált aggregátumokat (ez az SVM-ek bemenete, természetesen a másik megtanulandó paraméter a bemenő mintákból kikerült), egy, a kérdéses rekordokat tartalmazó (erre kell a címkéket előállítani) valamint három további táblát olvas be CSV fájlból az SVM tanítására és az eredmény ellenőrzésére. A processzelemek egy jelentős része azt a célt szolgálja, hogy a 3.2. részben leírtaknak a kimenet megfeleljen. A processz ábrája annak bonyolultsága miatt nem került a dolgozatba, mivel a RapidMiner processzek felépítésének és az alkalmazás működésének bemutatása túlmutat a TDK dolgozat keretein.

Az aggregátumokat kétszer-kétféleképp készítettem el. A verziók egyik dimenziója, hogy az SVM-et a tanítóhalmaz értékein tanítottam, majd a kiértékelő halmazon teszteltem, míg az esetek másik felében az eggyel korábbi vásárlásokon tanítottam és mind a tanítóhalmazon, mind pedig a kiértékelő adathalmazon teszteltem. A másik dimenziót, ami mentén két alverzió keletkezett, a hétfők, keddek, stb közötti hasonlóság indokolta. Az előző dimenzió szerinti mindkét algoritmustípusból kétféle keletkezett azáltal, hogy a dátumbecslést felbontottam-e a hét illetve a hét napjának becslésére. Ezutóbbi esetben a RapidMiner processz is változott, hiszen ekkor már három paramétert kellett becsülni a dátum kettébontása miatt.

## 3.5. Multi-label

*Multi-label* megoldást többféleképpen előállíthatunk a statisztikai módszerekből. A korábbi megoldásomhoz további eloszlásokat vettem fel, amelyek már a nap predikcióból indulnak ki. Ezeket az eloszlásokat szintén súlyoztam, majd a végleges tippbe beszámítottam. Az alábbi eloszlások figyelembevételét tettem lehetővé a döntésnél:

- felhasználónként a hét minden napján a 10 dolláronkénti lépcsőhöz egy statisztika.
- felhasználónként minden lépésmérethez (két vásárlás közt eltelt idő) egy statisztika pénzösszegekről

Ezeket az eredetileg relatív gyakoriság alapján előállított valószínűségeket a korábbiakhoz hasonlóan az idő szerinti súlyozással torzítottam, ezáltal az időben közelebbi eseményeket a következtetés szempontjából relevánsabb pozícióba állítottam. Többféle súlyozást és összevonást is kipróbáltam, ezekről részletesen a következő fejezetben esik szó.

Az adatbányászati megoldásaim esetében is elkészíthető *multi-label* változat. Ez a korábbi, 2.1. részben kifejtettek miatt azt jelenti, hogy a pénzösszeg becsléséhez felhasználom a dátumpredikciót is. Ezek alapján a korábbi RapidMiner processz is módosult. A változtatás lényege, hogy immár a következő alkalom időpontja bekerül a tanítómintába, majd az összegpredikcióknál felhasználásra kerül a dátumpredikció.

## 3.6. Vegyes megoldások

Több algoritmus esetében is megpóbáltam a megoldásokat ötvözni. Ennek a legegyszerűbb módja, ha az egymástól független dátum- és pénzösszegbecsléseket minden lehetséges kombinációban kipróbálom. Továbbá lehetőség van arra is, hogy az adatbányászati megoldásokhoz előállított aggregátumokhoz hozzávegyem a statiszikai módszer eloszlásközelítései közül azokat, amelyek felhasználónkénti plusz adatokat jelentenek. Egy másik lehetőség, hogy a *multi-label* módszerek közül azokat, amelyek az egyik már meglévő becslést használják fel a másik érték becslésére, már eleve egy másik módszer becsléseiből indulnak ki, illetve azzal együtt kerülnek kiértékelésre. A fennmaradó kombinációk, amelyek esetében az összefüggő attribútumkezelést másik predikción végezzük, mint a kiértékelést, szintén adnak egy opcióhalmazt, azonban ezek esetleg jól teljesítése elméleti síkon nehezen indokolható, hiszen ekkor az történik, hogy a rossz becslésekből következtetve jobb eredményt érünk el, mint a találatokból következtetve.

# Negyedik fejezet

# Eredmények és értékelés

A fejezetrész célja, hogy bemutassa az általam elkészített algoritmusokat, összehasonlítsa azokat kategórián belül és a legjobbakat kategóriák között. Célom továbba, hogy bemutassam a feladat és megoldásaim további lehetőségeit.

## 4.1. Kiértékelés

A kiértékelés az alábbi mutatók segítségével történik:

- a feladatkiírásnak megfelelő találati arány a teszt halmazon ($d_{\text{eff}}$)
- a helyezés, amelyet az előző érték alapján kaptunk volna (place)
- a feladatkiírásnak megfelelő találati arány a tanító halmazon ($d_{\text{eval}}$)
- dátum találati arány a tanító halmazon ($d_{\text{date}}$)
- összeg találati arány a tanító halmazon ($d_{\text{spend}}$)
- összeg találati arány a dátum találatokon belül a tanító halmazon ($d_{\text{sofd}}$) ($d_{\text{eval}}/d_{\text{date}}$)
- dátum találati arány az összeg találatokon belül a tanító halmazon ($d_{\text{dofs}}$) ($d_{\text{eval}}/d_{\text{spend}}$)

Sajnos az online kiértékelő weboldal jelenleg nem teszi lehetővé jelenleg, hogy az utolsó négy mérőszámot is kinyerhessük és összehasonlításra használhassuk a kiértékelő halmaz esetében is. A háttérben kiszámolja ugyan a rendszer, de semmilyen formában nem teszi elérhetővé. A kaggle.com [2] fórumán kapott válaszok alapján a fejlesztés várható, de jelenleg nem élvez elsődleges prioritást. Emellett annak lassúsága és a fórumban szereplő kérés alapján az online kiértékelő rendszert csak a kategóriánkénti legjobbnak tűnő algoritmusok esetében használtam. A csapatban beadott megoldás eredménye a fenti mérőszámok tekintetében látható a 4.1. táblázatban.

A továbbiakban minden módszerkategóriában ismertetek több, általam készített eljárást is. Az egyes kategóriákból a legsikeresebbeket kiemelem, majd összehasonlítom őket egymással, valamint a csapat által beadott megoldással is.

| $d_{\text{eff}}$ | place | $d_{\text{eval}}$ | $d_{\text{date}}$ | $d_{\text{spend}}$ | $d_{\text{sofd}}$ | $d_{\text{dofs}}$ |
|---|---|---|---|---|---|---|
| 16.26% | 48 | 16.26% | 40.11% | 33.90% | 40.54% | 47.96% |

**4.1. táblázat.** *A csapatban beadott megoldás mérőszámai*

## 4.2. Statisztikai megoldás

A következőkben az elkészített statisztikai megoldásaim eredményeit mutatom be. A megoldások abban különböztek, hogy más-más eloszlások alapján döntöttek, más súlyozást használtak az egyes közelítő eloszlások megalkotásakor, illetve, hogy ezek alapján máshogyan választották ki a megfelelő becslést. A továbbiakban a legjobban teljesítő megoldásokat mutatom be.


**4.1. ábra.** *A statisztikai megközelítésnél alkalmazott eloszlások, azok súlyozása, valamint a zajcsökkentést megcélzó módosítás*

Statisztikai megoldásaim közül a 4.1. ábrán bemutatottak szemléltetik jól az általam kipróbált és valamelyest sikeresnek talált algoritmusok körét, illetve az algoritmus felépítését. Jól látható, hogy mely eloszlásokat használja fel a statisztikai megoldásom, és azokat milyen súlyozás mentén veszi figyelembe. A „nap" illetve „nap$^2$" jelölések jelentése, hogy az eloszlásokat a relatív gyakoriságok nap sorszám illetve annak négyzete szerinti súlyozással számolom. A nap sorszám 2010. április 1-jén 1, onnantól naponta eggyel nő. Az ábrán legvilágosabb kék színnel szerepel a naptípusonkénti eloszlás, mert az a kimeneti rangsorba négyzetesen számít, emellett a két lépésméret eloszlást összevonva jelöltem, mivel a második lépésméret valószínűség nagyon gyakran nulla a megoldásoknál is, emiatt a végső megoldásban a rangsorolásba a két lépésméret-valószínűség összege számít be. Az aktuálisan „legjobb" megoldás végül úgy állt elő, hogy a zajcsökkentési célokat szolgáló ünnepnap kiszűrés is ki lett kapcsolva, mert az eloszláskombinációk és a súlyozás miatt többet ront, ha hiányzik két hétnyi adat, mint amennyit a zaj eldobásával nyerünk.

Az eloszlások kummulálására többféle módszert is kipróbáltam, de a legjobbnak az a módszer bizonyult, amikor egyes „valószínűségek" szorzata alapján állítottam fel a sorrendet, a fontosabb paramétereket a szorzatban hatványozottan szerepeltetve. Több esetben előfordult, hogy olyan eloszlás is bekerült a rangsorolásba, amely nagyon sok esetben nulla valószínűséget ad, emiatt

a szorzatba csak a vele leginkább rokon eloszlás valószínűségértékéhez hozzáadva került bele. A rangsorolásnál a legjobban teljesítő megoldás a három kékkel jelölt valószínűség szorzata alapján választja ki a medián dátumot. A pénzösszeg esetében pedig – mivel egyetlen eloszlás alapján kell döntést hozni – egyszerűen a módusz szerint döntünk. A kiértékelésnél (4.2. táblázat) négy változat összehasonlítása szerepel, ezek rövid leírása az alábbi:

| | |
|---|---|
| 1 | a 4.1. ábra szerinti algoritmus |
| 2 | 1, a kétlépéses eloszlás használata nélkül |
| 3 | 1, ünnepnapok elvétele nélkül |
| 4 | 1, a naptípus szerinti eloszlást egyszeresen véve |

| **sorszám** | 1 | 2 | **3** | 4 |
|---|---|---|---|---|
| $d_{\text{eval}}$ | 15.64% | 15.45% | 15.75% | 15.53% |
| $d_{\text{date}}$ | 41.64% | 41.02% | 41.75% | 40.81% |
| $d_{\text{spend}}$ | 35.10% | 35.10% | 35.10% | 35.10% |
| $d_{\text{sofd}}$ | 37.56% | 37.66% | 37.72% | 38.05% |
| $d_{\text{dofs}}$ | 44.56% | 44.02% | 44.71% | 44.25% |

**4.2. táblázat.** *A statisztikai megoldásaim futáseredményei*

Ahogy az látható, a 3-assal jelölt megoldás szinte minden tekintetben megveri a többit. Emiatt is került a végleges megoldásból az ünnepnapok szűrése. Ez a megoldás az online ellenőrzőrendszerben 15.64%-ot ért el, amely a 62. helyre elegendő. Tehát összességében még gyengébb, mint a csapatmegoldás, azonban mindkét részkategóriában (dátum és pénzösszeg) jobban teljesít, tehát a dátumok felhasználása az összegpredikcióban segíthet.

## 4.3. Adatbányászati megközelítés

Az adatbányászati megoldások között a különbséget az aggregátumok, valamint a használt tanulóeljárás határozta meg. Mind futási sebességben, mind pedig a mérőszámok tekintetében alulmaradtak a korábban bemutatott statisztikai módszerekkel szemben. Az elkészített adatbányászati megoldásokat négy kategóriába soroltam korábban, ez látható a 4.3. táblázatban. A csoportok közül minden tekintetben a legjobb a III. jelöléssel ellátott csoport teljesítménye volt, illetve kevéssel elmaradva a IV.-es csoport algoritmusai teljesítettek még viszonylag jól. Az I. illetve II. kategória algoritmusai alig több, mint fele akkora hatékonysággal tudtak becslést adni. Mindezek miatt a futáseredmények összehasonlításában csak a III.-as kategória algoritmusainak egy részét mutatom be.

| **tanító minta ↓** | **dátumbecslés egyben** | **hét és hét napjának becslése külön** |
|---|---|---|
| **tanító adathalmaz** | I | II |
| **korábbi vásárlás** | III | IV |

**4.3. táblázat.** *Az adatbányászati megoldásaim csoportosítása*

Az általam elkészített, a kategóriájukban legjobbnak bizonyult adatbányászati módszerek jellemzői (mind a III. algoritmuscsoportból kerültek ki) a 4.4. táblázatban szerepelnek a kiértékelés eredményével, azaz a mutatószámokkal együtt. Látható, hogy a legjobbnak bizonyuló megoldás is nagyon elmaradt a statisztikai megoldások teljesítménye mögött. Ez könnyen magyarázható, hiszen nagyon sok olyan információ van, amelyet a statisztikai módszerek esetében lehetőség volt felhasználni, míg itt a jelenlegi struktúra ezt nem teszi lehetővé. Ilyen tényező például a következő vásárlások eloszlása, valamint – annak ellenére, hogy egy minimum oszlopot az aggregátumtáblába felvettem – az sem garantált, hogy a predikcionál a 2011. április elseje előtti napok nem lesznek figyelembe véve. Látható az is, hogy a tanulóiterációk közel duplájára növelése nagyon kis plusz nyereséget okoz, mégegyszer megduplázva pedig már az 1500 iterációnál látott értékeket kaptam. Emiatt pusztán az iterációk növelése igazoltan nem alkalmas az algoritmusok hatékonyságának javítására, emellett a futásidőt is megnöveli.

| dátumbecslő SVM típusa | dot | dot | radial (RBF) | dot |
|---|---|---|---|---|
| összegbecslő SVM típusa | dot | radial (RBF) | radial (RBF) | radial (RBF) |
| tanulóiterációk száma | 800 | 800 | 800 | 1500 |
| $d_{\text{eval}}$ | 6.40% | 10.99% | 10.99% | 11.00% |
| $d_{\text{date}}$ | 33.33% | 33.33% | 33.33% | 33.34% |
| $d_{\text{spend}}$ | 20.43% | 27.82% | 28.82% | 27.87% |
| $d_{\text{sofd}}$ | 19.20% | 32.97% | 32.97% | 32.99% |
| $d_{\text{dofs}}$ | 31.33% | 39.50% | 39.50% | 39.47% |

**4.4. táblázat.** *Az adatbányászati megoldásaim futáseredményei*

Sajnos, mint a mérőszámok is mutatják ez a megoldás jelenlegi formájában kevés sikert hozott. Ennek oka, hogy kevés a bemeneti adat egy-egy vásárlóra nézve, illetve kevésnek bizonyult az egyetlen közös modell is. Azonban a bemeneti adatok és a modellek számának megnövelése nem kívánt futásidőnövekedéssel jár.

## 4.4. Multi-label megoldás

A feladaton az általam elkészített megoldások közül a *multi-label* megoldások teljesítettek a legjobban. Ezek közül is elsősorban a tisztán statisztikai megközelítések, amely már a 3.5. részben részben bemutatásra került. Az algoritmus működéséhez felhasznált eloszlások és azoknak megfelelő súlyozás a statisztikai módszereknél leírtakhoz hasonló jelölésrendszer szerint szerepel a 4.2. ábrán. A világosabb színnel jelölt naptípusonkénti valószínűség szintén négyzetesen szerepel a szorzatban, emellett a másik két eloszlás szerinti valószínűség összege szerepel a szorzatban. A 4.5. táblázatban szerepelnek az egyes többcímkés megoldások futáseredményei. A mérési adatokkal együtt bemutatott eljárások alapelvének rövid leírása az alábbi:


**4.2. ábra.** *A multi-label megközelítésnél alkalmazott eloszlások és azok „torzító" súlyozása*

| | |
|---|---|
| ML1 | A 3.5. részben bemutatott és a 4.2. ábrának megfelelő megoldás |
| ML2 | ML1, a lépésméretnek megfelelő eloszlás nélkül |
| ML3 | ML1, a naptípusnak megfelelő eloszlást csak egyszeresen véve |
| ML4 | ML1, a naptípusnak megfelelő eloszlás háromszorosan véve |

| **sorszám** | **ML1** | ML2 | ML3 | ML4 |
|---|---|---|---|---|
| $d_{\text{eval}}$ | 17.61% | 17.44% | 17.58% | 17.56% |
| $d_{\text{date}}$ | 41.75% | 41.75% | 41.75% | 41.75% |
| $d_{\text{spend}}$ | 35.57% | 35.46% | 35.93% | 35.31% |
| $d_{\text{sofd}}$ | 42.18% | 41.77% | 42.11% | 42.06% |
| $d_{\text{dofs}}$ | 49.51% | 49.18% | 48.93% | 49.79% |

**4.5. táblázat.** *Multi-label megoldásaim futáseredményei*

Látható, hogy az elkészített *multi-label* megoldások hatásfoka egymással milyen viszonyban áll. Érdekesség, hogy míg a külön-külön sikeresség dátum és pénz tekintetében kisebb az ML1 esetében, mint az ML3-nál, mégis a metszet mérete nagyobb. Az online értékelő rendszerben 17.74%-os eredményt ért el az ML1 predikciója, ez a 18. helyre lett volna elegendő.

Mivel az SVM önmagában is rosszul teljesített, ezért a kiértékelésben sem szerepel *multi-label* változata, amikor is a dátumot is bevezetjük az összegbecslés tanítóbemenetére és a dátumpredikciót pedig hozzáadjuk a modellhez. Az SVM teljesítménye ezzel a módosítással valamelyest nőt, de nem hozott látható áttörést.

## 4.5. Több megoldástípus ötvözése és összehasonlítás

Több, különböző kategóriából vett, a kategórián belül jól teljesítő megoldástípus ötvözése esetén nem fordult elő olyan, hogy az előálló megoldás mindkét eredeti módszernél jobban teljesített volna. Ez köszönhető annak is, hogy az adatbányászati megoldások sokkal gyengébbek, mint a

statisztikai illetve *multi-label* algoritmusok. Emellett valamelyest a véletlennek is köszönhető, hogy a legjobb *multi-label* algoritmus és a csapatban beadott megoldás ötvözete nem teljesített jobban. Mindkét lehetséges párosítás jobban teljesített, mint a gyengébb, de rosszabbul, mint a *multi-label* algoritmus. Ebből azt a következtetést is levonhatjuk, hogy a csapatban beadott megoldást sikerült minden tekintetben megverni, ezzel az alapjában véve másfajta megközelítéssel.

Kategóriánként egy-egy megoldás értéket tartalmazza a 4.6. táblázat. Mindegyik algoritmusnál a kiértékelésnél kapott százalékos érték szerepel és az a helyezés, amelyet azzal elérhettünk volna a verseny határidejének lejárta előtt. Az induló 297 csapatból tehát lehettünk volna akár 18. helyezettek is.

| | **statisztikai** | **SVM** | **csapatban beadott** | ***multi-label*** |
|---|---|---|---|---|
| $d_{\text{eff}}$ - hatásfok | 15.64% | 11.17% | 16.26% | 17.74% |
| place - elérhető helyezés | 62 | 180 | 48 | **18** |

**4.6. táblázat.** *Kategóriánként legjobb algoritmusaim online tesztelési eredményei*

## 4.6. Fejlesztési lehetőségek és összegzés

A látszólag rosszul teljesítő adatbányászati módszereket lehetőség van jelentősen javítani azzal, ha – ahogyan a csapat által beadott megoldásban is tettük a $k$-NN algoritmust használva – a modelleket nem, vagy nem önmagában a közös aggregátumok alapján készítem el, hanem a használt ablakot a teljes múltbéli adathalmazon végigtolom 2011. április 1-ig. Ha ezeket a tapasztalt „következő vásárlásokat" használnám fel a regressziós feladat tanítására, akkor különböző modellt tudnék alkotni minden egyes vásárlóhoz. Sajnos ennek a módszernek egy igen nagy hátulütője, hogy a tanítási fázis nagyon hosszú, órákban illetve napokban mérhető, ezért a TDK dolgozatban ennek a módszernek a kezdetleges változata sem kapott helyet. Ezzel a módosítással jobb eredményeket érhettem volna el az adatbányászati algoritmusokkal is, de csak nagyon időigényes tesztfuttatásokkal. Ezt alátámasztja az is, hogy a $k$-NN is közel ugyanilyen rosszul teljesített egyetlen modell esetén, míg a csapattal készített megoldásban a vevőnkénti modell sokkal jobb eredményeket hozott.

A jelenleg legjobban teljesítő megoldás hatalmas előnye, hogy bármikor továbbfejleszthető újabb eloszlás felvételével vagy a súlyozások jobb eredményt hozóra cserélésével. Mindkét esetben több lehetőség is rendelkezésre áll. Akár mindkét kategóriába besorolható, ha felveszünk közelítő eloszlásokat a pénzösszegekre az eltelt napok száma alapján úgy is, hogy minden egyes vásárló esetében próbáljuk regresszió segítségével megállapítani, hogy melyek a legvalószínűbb vásárlási értékek. Igen sok lehetőség rejtőzhet a vásárlónkénti segédmodellek alkalmazásában, amellyel az emberek között különbségekből következő pontatlanságok egy részét lennénk képesek kiküszöbölni. Nagy előnye a jelenlegi megoldásnak mindezeken túl, hogy mivel eloszlásokat készítettem az elkészült predikcióhoz mindig adott egy konfidencia-érték (hogy ez ténylegesen 0 és 1 közötti

érték legyen normálni kell, a valószínűségek esetenkénti összeadása miatt, vagy az összeadásoknál be kell vezetni az 1-et, mint felső korlátot) is, amely alapján lehetséges a bizonytalanul predikált halmazt leválogatni és azon más algoritmust, algoritmusokat alkalmazni. Mindezeken felül még a pénzösszegbecslés lépésmérete is tovább csökkenthető egy pontosabb eloszlás érdekében.

Összefoglalva tehát elmondható, hogy egy jó hatékonyságú többcímkés algoritmus készült el a dolgozat keretei között, amely még többféleképpen tovább javítható, moduláris, emellett a predikciók mellé egy konfidenciaértéket is biztosít, tehát már előre elkülöníthetők a valószínűleg rosszul osztályozott rekordok. A megoldással a versenyben is előkelő, 18. helyen végeztem volna. Ezek alapján elmondható, hogy jelen feladat esetében a *multi-label* megközelítés használata kifizetődő volt.

# Köszönetnyilvánítás

Köszönettel tartozom Gáspár-Papanek Csabának, konzulensemnek és Nagy Gábornak a DM-Lab munkatársának, amiért a feladattal megismertettek és a csapatversenyben részt vettek. Köszönöm türelmüket és tanácsaikat. Szintén köszönet illeti a SZIT tanszék ZH-javító csapatát, amiért a TDK dolgozat készítésének idején megpróbáltak valamelyest tehermentesíteni. Köszönöm ezenfelül a türelmet mindazoknak, akik a dolgozat készülése közben ezzel adóztak felém: kedvesemnek, családomnak, lakótársaimnak és a Nagytétényi Úti Kollégium több lakójának.

# Ábrák jegyzéke

1.1. a megelőző egy év vásárlásaikor kifezetett pénzösszegek közül ennyi esik egy-egy érték tízdolláros környezetébe (a tanító és a kiértékelő adathalmazon együttesen) – a grafikon értelmezési tartománya az ábrán láthatónál bővebb, a nagyobb értékek szemléltetési célból hiányoznak . . . . . 7
1.2. a megelőző egy év egyes napjaihoz tartozó elköltött összegek átlagértéke (összehasonlítva a tanító és a kiértékelő adathalmazon) . . . . . 7
1.3. a megelőző egy év egyes napjaihoz tartozó elköltött összegek mediánja (összehasonlítva a tanító és a kiértékelő adathalmazon) . . . . . 8
1.4. a megelőző egy év egyes napjaihoz tartozó vásárlások száma (a tanító és a kiértékelő adathalmazon együttesen) . . . . . 8
1.5. a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó vásárlások száma (a tanító adathalmaz adatai alapján) . . . . . 8
1.6. a megelőző egy év egyes napjaihoz tartozó vásárlások átlagértéke és mediánja (a tanító és a kiértékelő adathalmazon együttesen) . . . . . 9
1.7. a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó vásárlások átlagértéke és mediánja (a tanító adathalmaz adatai alapján) . . . . . 9
1.8. a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó első vásárlások száma (a tanító adathalmaz adatai alapján) . . . . . 10
1.9. a következő nyolcvan nap (a kérdéses intervallum) egyes napjaihoz tartozó első vásárlásokhoz tartozó pénzösszegek átlaga és mediánja (a tanító adathalmaz adatai alapján) . . . . . 10

3.1. az általam készített statisztikai megközelítés dátumbecslésének döntéshozatalához szükséges információk előállítása – a relatív gyakoriságokat vásárlónként és összességében is előállítottam . . . . . 17

4.1. A statisztikai megközelítésnél alkalmazott eloszlások, azok súlyozása, valamint a zajcsökkentést megcélző módosítás . . . . . 22
4.2. A *multi-label* megközelítésnél alkalmazott eloszlások és azok „torzító" súlyozása . 25

# Táblázatok jegyzéke

1.1. A rendelkezésre álló adatok mennyiségi jellemzői . . . . . 6
1.2. A rendelkezésre álló tranzakciótörténet részletek időhatárai . . . . . 6

4.1. A csapatban beadott megoldás mérőszámai . . . . . 22
4.2. A statisztikai megoldásaim futáseredményei . . . . . 23
4.3. Az adatbányászati megoldásaim csoportosítása . . . . . 23
4.4. Az adatbányászati megoldásaim futáseredményei . . . . . 24
4.5. *Multi-label* megoldásaim futáseredményei . . . . . 25
4.6. Kategóriánként legjobb algoritmusaim online tesztelési eredményei . . . . . 26

# Irodalomjegyzék


[1] Hellinger Péter. Logitboost alapú osztályozó eljárás működésének vizsgálata. Master's thesis, Budapesti Műszaki és Gazdaságtudományi Egyetem, 2009.

[2] Dunnhumby UK and kaggle.com. dunnhumby's shopper challenge. http://www.kaggle.com/c/dunnhumbychallenge, 2011.10.27.

[3] Rapid-I. weboldal. http://rapid-i.com, 2011.10.27.

[4] TEXnicCenter. weboldal. http://www.texniccenter.org, 2011.10.27.

[5] MikTEX. weboldal. http://www.miktex.org, 2011.10.27.

[6] Tikk Domonkos. *Szövegbányászat.* Typotex Kft., 2007.

[7] Zhi-Hua Zhou Min-Ling Zhang. Ml-knn: A lazy learning approach to multi-label learning. *National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China*, 2006.

[8] I. Vlahavas I. Katakis, G. Tsoumakas. Multilabel text classification for automated tag suggestion. *Proceedings of the ECML/PKDD 2008 Discovery Challenge, Antwerp, Belgium*, 2008.

[9] Tao Mei Jingdong Wang Guo-Jun Qi Zengfu Wang Zheng-Jun Zha, Xian-Sheng Hua. Joint multi-label multi-instance learning for image classification. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference, Anchorage, AK*, 2008.

[10] G. Kalliris I. Vlahavas K. Trohidis, G. Tsoumakas. Multilabel classification of music into emotions. *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 325-330, Philadelphia, PA, USA*, 2008.

[11] Olga G. Troyanskaya Zafer Barutcuoglu, Robert E. Schapire. Hierarchical multi-label prediction of gene function. *Bioinformatics 22 (7)*, 2006.

[12] Gareth Funka-Lea Leo Grady. Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials. *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*, 2004.

[13] I. Vlahavas G. Tsoumakas. Random k-labelsets: An ensemble method for multilabel classification. *Proc. 18th European Conference on Machine Learning (ECML 2007), Warsaw, Poland*, 2007.

[14] Grzegorz Swirszcz Vikas Sindhwani Yan Liu-Prem Melville Dong Wang Jing Xiao Jianying Hu Moninder Singh Wei Xiong Shang Yan Feng Zhu Alexandru Niculescu-Mizil, Claudia Perlich, editor. *Winning the KDD Cup Orange Challenge with Ensemble Selection*, 2009.

[15] Tibshirani R. Friedman J., Hastie T. *The elements of statistical learning Data mining, inference, and prediction - 2ed.* Springer, 2008.

[16] Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*, 1956.

[17] Parzen E. On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*, 1962.

[18] D'Yakonov Alexander. My method for dunnhumby's shopper challenge problem solving. competition winner, 2011.

[19] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[20] Pataki B. Strausz Gy. Takács G. Valyon J. Altrichter M., Horváth G. *Neurális hálózatok.* Panem Kft., 2006.