10.24387/CI.2017.1.3

Oláh Judit – Erdei Edina – Popp József

Értékesítési adatok klaszteranalízise és előrejelezések készítése SAP HANA Platformon

A versenyképesség szükséges feltétele, illetve növelésének egyik lehetséges módja, ha megteremtjük a vállalati működést teljes mértékben átfogó, központosított, könnyen áttekinthető informatikai hátteret. A cikk szerzőinek kutatási célkitűzése egy kereskedelemmel foglalkozó vállalat értékesítési tranzakcióit az SAP HANA, illetve SAP Predictive Analytics által biztosított lehetőségekkel elemezni a vállalat készletezési, értékesítési, marketing stratégiájának javítása érdekében.

Bevezetés

Napjainkban egyre több kis- és középvállalat működteti, illetve felügyeli pénzügyi, logisztikai, termelési, humán-erőforrás és egyéb tevékenységeit Enterprise Resource Planning (ERP) informatikai rendszerekkel, melyek az említett folyamatokat egységes keretben képesek kezelni. Adataink mélyebb elemzéséhez az adattáblákat felépítő oszlopokat, attribútumokat összességükben indokolt kezelni, azaz olyan módszerekre van szükség, amelyek ezeket a dimenziókat együttesen elemzik. Ilyen elemzési eszközök a legjobb, naprakész adatbányászati algoritmusok, amelyeket a Big Data és Data Science területén az elmúlt évtizedben a legeredményesebben alkalmaztak. A kutatásunk célja olyan elemzési és előrejelzési módszerek bemutatása, amelyek vállalatok különböző pénzügyi és logisztikai problémáira nyújt megoldást.

A célokhoz rendelt legfontosabb feladatok közé tartozik az adathalmaz különböző relációkból történő összeállítása a HANA adatbázis-kezelőn belül. A következő lépés az adatok hónapokra történő aggregálása és egy 3 dimenzióból álló kocka kialakítása. Ezen adathalmazt negyedévre bontottuk fel, majd a különböző időszakokra a Predictive Analytics, valamin az R statisztikai programnyelv használatával K-közép klaszterezési eljárást alkalmaztunk. Elemzéseink másik alappillére a 15 negyedévre készített klaszterezési kimenetekre alapozott idősoranalízis (2013. 01. 01. – 2016. 10. 01.), amely segítségével a 15 időszakra kapott klaszterközéppontok időbeni alakulására készítettünk előrejelzéseket. A klaszteranalízis során olyan javaslatokat tettünk, amelyek a stratégia javításával a vállalkozás versenyképességét, valamint üzleti tevékenységének eredményességét növelik. Idősor-analízis segítségével a vállalatra vonatkozóan olyan előrejelzéseket készítettünk, amelyek megkönnyítik a vizsgált vállalat készletgazdálkodási és árazási tevékenységét.

Szakirodalmi áttekintés

Az SAP HANA kialakulása

Az 1972-ben megalapított SAP AG (System, Applications & Products in Data Processing) egyike azon vezető vállalatoknak, amelyek integrált vállalatirányítási rendszereket gyártanak. A fő vállalatirányítási termékei közé sorolhatjuk az SAP Business One-t, mely a kis- és középvállalatoknak nyújt segítséget üzleti folyamataik egyszerűbb kezeléséhez (WOLFGANG, 2009). Az SAP által forgalmazott és fejlesztett, memória-alapú, oszlop-orientált, HANA relációs adatbázis-kezelő rendszer 2010-ben került piacra. A felhasználók ekkor úgy vélték, hogy az alkalmazás még nagyon kiforratlan állapotban van, ezért a piaci bevezetést követő 3 évben lényeges fejlesztésen ment keresztül. Az alkalmazás megjelenéséhez rendkívül fontos volt a több terrabájt kapacitással rendelkező rendszermemória használata, melynek köszönhetően az adatbázis memória-alapúvá vált. A számítógépes adattárolás fő memóriájára épül, így nagyobb teljesítményt nyújt, mint a lemezes tároló-mechanizmust alkalmazó adatbázis-kezelő rendszerek (MARK, 2013).

A HANA-hoz akár mobil eszközök segítségével is kapcsolódhatunk, ami hatékonyabb munkavégzést eredményezhet, ezáltal pozitív hatással lehet a versenyképességre is. Ily módon a nap minden percében a legfrissebb információk birtokában hozhatnak döntéseket a menedzserek (PENNY et al., 2015).

Az új technológiára olyan erőforrás, ami a meglévő rendszerekben rejlő információk hatékony kiaknázására tökéletesen alkalmas, éppen ezért a hangsúly a nagyméretű adatbázisokból gyorsan készíthető riportokon, illetve átfogó adatelemzéseken van.

In-memory technológia

A HANA a kis- és középvállalatok esetén egy független B1A (Business One Analytics) szerveren fut, ahol a teljes SAP Business One adatbázist betölti a memóriába, és a lekérdezéseket hajtja végre. Az in-memory technológia segítségével az adatok különféle nehézségek (pl. kapacitáshiány) nélkül, pillanatok alatt elérhetővé válnak, következésképpen nincs szükség az adatok elő-feldolgozására (preprocessing), csoportosítására, ami fáradtságos és időigényes munka lenne (HASSO, 2012). A memóriába a ,,hot data" adatok kerülnek, avagy azok az adatok, amelyek elérésére gyakran van szükségünk. A fő problémát a következő példa szemlélteti: ,,A HANA egy DNS-analízissel 1 perc alatt végez, míg mindez egy klasszikus, háttértár-alapú rendszer esetében nagyjából 2 napot vesz igénybe".

A memória-alapúságnak köszönhetően az analitikai lekérdezések segítségével a felhasználók sokkal gyorsabban hozhatnak a vállalat irányítására vonatkozó megalapozottabb döntéseket. A kimutatások alapján a vállalatok a veszteséges termékeket kivonhatják piacaikról és az eseményekre történő azonnali reagálás következtében a profitot is könnyebben maximalizálhatják.

Oszlop-orientált adatbázisok előnyei

Az adatok tárolása az SAP Business One rendszerben sem a már régen megszokott MS SQL szerveren történik – ahol soros adatszervezés van jelen –, hanem egy sokkal eredményesebb oszlop-orientált konstrukción. Ez a formátum jobban illeszkedik az elemzésekhez, ahol sokszor oszlopokban, homogén adatokkal algoritmikus műveleteket (sorba rendezés, összegzés, szűrés, átlagolás) szükséges elvégezni. Gyakran az adatbázis oszlopain hajtunk végre matematikai eljárásokat, melyhez az adatokat könnyen be tudjuk olvasni, míg a soralapú technológiáknak ehhez az egész táblát fel kell dolgozniuk (NIELS et al., 2012). Használatával a korábban akár több tíz percet igénybe vevő riportok elkészítése másodpercek alatt megtörténik és az ügyviteli rendszert sem terhelik le. Míg a soralapú adatbázisban egy új rekord felvételéhez a meglévő rekord indexeinek frissítésére van szükség, addig az oszlopalapú adatbázisban nincs szükség analitikai indexekre, így a frissítésre fordított idő elhanyagolható. Az indexek bővülése és karbantartása nehézségeket okozhat a soralapú rendszerekben (SAP SE, 2014). Az oszlopalapú implementációnak köszönhetően sokkal inkább indokolt és kivitelezhető a párhuzamos lekérdezés, ennek segítségével a teljesítmény is tovább javítható.

OLAP rendszerek

Az online analitikus feldolgozás (On-Line Analytical Processing) az adatbázis-kezelő rendszerek számára lehetővé teszi, hogy a felhasználók nagyon gyorsan le tudják kérdezni a számukra fontos adatokat. Az OLAP olyan adatbázis-technológia, amelyet tranzakciók végrehajtása helyett lekérdezések és kimutatások használatára optimalizáltak (OMAR et al., 2015).

A technológia az 1970-es évek során kezdett gondolkodásba ejteni sok fejlesztőt, amikor megpróbáltak egy rugalmas, felhasználóbarát kezelőfelületet kialakítani a szervezetek vezetői számára, hogy azok stratégiai döntéseiket hatékonyan és biztonságosan tudják megvalósítani. A megfelelő hardver, valamint szoftvertámogatással az elemző, analitikai alkalmazások az 1990-es évek elejére fejlődtek ki (AHSAN, 2009).

Az OLAP 1994-ben jelent meg, amely teljesen új lendületet vitt a döntéstámogató rendszerek fejlődésébe, hiszen az első kereskedelmi célú, web-orientált rendszerek ennek köszönhetően a következő években bukkantak fel, majd kerültek piacra. Érdemes megemlíteni, hogy magát az OLAP fogalmát E. F. Codd vezette be a hétköznapokba, mely mára már általánosan elfogadottá vált. Az utóbbi években robbanásszerű növekedés volt megfigyelhető nemcsak a kínált termékek és szolgáltatások számában, hanem ezeknek a technológiáknak az iparban történő alkalmazásában is. Az OLAP magába foglalja a többdimenziós modellekbe rendeződő adatok iteratív feldolgozását, lekérdezését, vagyis ezek az elemzések információellátó feladatként szolgálnak, melyet a rendszer nagy teljesítményével hatékonyan képes feldolgozni. A rendszer egyik legfontosabb jellemzője, hogy akár egyidejűleg többen is használhatják, azaz megosztott adatforrással dolgozik. Ez a megosztás több különböző, inhomogén adatforrást rejt magában, ami azt jelenti, hogy a műveletekbe bevont adatok több különböző adatbáziskezelőből kerülhetnek beolvasásra.

Az OLAP technológiát az egyszerűbb helyett az összetettebb lekérdezések, komplexebb számításokat magukba foglaló utasítások, nagyobb adatmennyiségek jellemzik. Különös gondot kell fordítani a hatékony válaszgenerálási algoritmusokra, továbbá figyelmet igényel, hogy a lehetőségekhez mérten minél rövidebb idő alatt hajtsa végre a rendszer a felhasználó által kiosztott feladatokat. Ezen cél eléréséhez például a rendszer előre letárolja a különböző előszámítások, részműveletek eredményeit (HELEN – ANINDYA, 2001).

Az adatbázis a modellezett rendszerben található egyedeknek az éppen érvényes állapota, értéke mellett a múltbeli adatait, vagyis azok történetét is tárolja. Egy vállalati rendelési információs rendszer esetében például az élő rendelések mellett a korábbi időszakok rendeléseit is tárolja. A múltbeli adatok felhasználásával pontosabb lehet az elemzési munka, a jövő előrejelzése.

OLAP műveletek

Az OLAP rendszer fő célja az operatív adatbázisból nagy adatmennyiséget érintő – a párhuzamosan futó műveletekkel történő – gyors adatkinyerés. Az adatokat többdimenziós kockákban tároljuk, melynek élei azok a szempontok, amelyek alapján adatainkat összegezni, vagy elemezni szeretnénk, mezői pedig a számunkra releváns információk. Az OLAP-adatbázisok alapvetően kétféleképpen tekintenek az adatokra: a kockák celláiban található értékekre, amelyek elő-feldolgozhatók, aggregálhatók és elemezhetők, illetve dimenziókra.

A kockákból kinyerni kívánt adatokban rejlő információkat komplex algoritmusok segítségével valósíthatjuk meg. Az OLAP a modell-vezérelt analízist támogatja, szervezettsége hierarchikus, így a kimutatásokban nem jelent nehézséget az értékesítések magas szintű (például régiónkénti aggregálás) megjelenítése és a különösen alacsony/magas forgalmat lebonyolító telephelyek adatainak kiemelése (GARCIA et al., 2015).

Az OLAP alapvető művelete az adatkocka létrehozása. A tipikus OLAP-műveletek közé soroljuk még a következőket:

- Szeletelés (slicing): cellák egy olyan csoportjának a kiválasztását jelenti a teljes többdimenziós tömbből, amelyet értékeknek egy vagy több dimenzió menti rögzítésével kapunk.
- Kockázás (dicing): cellák egy olyan részhalmazát jelenti, amelyet attribútum-értékek egy tartományának megadásával kapunk. Ez ekvivalens azzal, hogy a teljes tömbből egy résztömböt választunk ki.
- Göngyölítés (roll-up): a hierarchia teszi lehetővé ezt a műveletet. Az eladási adatokat összegezhetjük például hosszabb időszakokra is. Dimenzión belül összesítjük az adatokat, nem pedig a dimenzió mentén.

1. számú táblázat: A vizsgált vállalat legfontosabb adatai

Jellemző	Érték
Eladási rekordok száma (db)	7 052 888
Vizsgált időszak kezdete	2013. 01. 01.
Vizsgált időszak vége	2016. 10. 01.
Értékesített termékek (db)	15 674
Jelenleg aktív termékek (db)	8 392

Forrás: saját szerkesztésű táblázat

- Lefúrás (drill down): egy olyan adattábla esetén, ahol az idő dimenzió hónapokra van bontva, a havi eladásokat bonthatjuk napi szintre.
- Forgatás (pivoting): átalakítja az adatok többdimenziós képét, az eredmény egy két dimenzióból álló kontingenciatáblázat lesz (például az időpont és a termék körüli forgatás) (MICHELANGELO et al., 2013).

Megvalósítási módszerek

A modern rendszerek olyan eszközöket nyújtanak, melyek segítségével elemezhetők az OLAP kockák adatai és a relációs adatbázisokban tárolt adatok is. Különbséget kell tennünk aközött, hogy milyen adatbázisban szeretnénk tárolni az adatokat. A napjainkig kialakított három megvalósítási módszer a következő (MOLAP, ROLAP, HOLAP):

- *Relational OLAP:* az adatok tárolását hagyományos relációs adatbázis-kezelővel végezzük. Elterjedtsége főképp rugalmasságára és a relációs adatbázis-kezelők viszonylag alacsony árára és megbízhatóságára vezethető vissza. A HANA adatbázisban a relációs tárolási technikát az „Analytical View" nézetben láthatjuk, ahol csillagsémát (star scheme) alkalmazva átláthatóbbá tettem a táblák közötti összefüggéseket, kapcsolatokat.
- *Multidimensional OLAP:* speciális adatbázis-kezelővel közvetlenül, valamely többdimenziós célstruktúrában (pl. tömbökben) tárolják az adatokat, és a MOLAP-szerverek ezekkel az adatokkal valósítják meg a műveleteket.
- *Hybrid OLAP:* az adatbázis-kezelő biztosítja a hagyományos relációs tárolás lehetőségei mellett a többdimenziós tárolási metódusokat is. Itt az előző kettőre jellemző adatbázis sémák mindegyike előfordulhat. Egyre inkább tendencia, hogy a relációs adatbázisok támogatják a többdimenziós adattárolást. Tulajdonképpen a HANA egy HOLAP adatbázis-kezelőnek tekinthető, mert mind a két – az adatok elemzéséhez szükséges – módszer megtalálható a kialakított felületek között (OKSANA et al., 2010).

A kutatási módszer

Az adatbázis bemutatása

Az SAP Business One vállalatirányítási rendszer 2008-ban vezette be a vizsgált vállalat. Az elemzések szempontjából legfontosabb a vállalatra vonatkozó információkat az **1. számú táblázat** tartalmazza.

Az eladásokat tartalmazó adathalmaz jelentős, tranzakciókkal kapcsolatos attribútumot is tartalmaz. A megfelelő elemzések, kimutatások, majd előrejelzések elvégzéséhez célszerű a vizsgálatot a következő attribútumokra szűkíteni: vevőkód (*CardCode*), cikk kód (*Itemcode*), eladás dátuma (kalkulált mező: hónapra összegzés), árrés (kalkulált mező), beszerzési érték (kalkulált mező), rendelt mennyiség (*Quantity*), eladási ár (*LineTotal*).

Az elemzés módszerei

A HANA elsősorban valós idejű riportok, analitikák generálására, illetve az adatbázisra történő lekérdezések futtatására alkalmas. Felületéhez kizárólag felhasználónévvel és jelszóval rendelkező, a rendszert üzemeltető és a kimutatásokat elkészítő felhasználó férhet hozzá. További módszereink közé tartoztak az alábbi eljárások: *K*-közép, valamint sűrűség alapú (DBSCAN) klaszterezési algoritmusok. Ezen klaszterezési módszerekkel megtaláltuk az értékesítési adatok, cikkek közötti összefüggéseket, majd az így felfedezett klaszterek, homogén csoportok középpontjainak koordinátáit jeleztük előre (idő és más egyéb változók függvényében) regressziószámítás segítségével. Mind a klaszterezési eljárások esetében, mind a regressziószámítás során segítségünkre voltak a Predictive Analytics operátorai, valamint az *R* statisztikai programnyelv. Ezen felül kiemelnénk a jól ismert döntési fákon alapuló (azokon belül is a C4.5 és CHAID algoritmusokra építő), naiv-Bayes és kNN osztályozási technikákat is, amiknek segítségével az adathalmaz alapján egy prediktív modellt készítettünk ismeretlen, új rekordok osztályozásához. Elemzési módszereink között találhatjuk a regresszióanalízist is, mely két vagy több véletlen változó között fennálló kapcsolat modellezésére szolgál.

A kutatási eredmények

A klaszterezés eredményei

Az SAP Predictive Analytics-ben a korábban (2013. 01. 01. – 2016. 10. 01.) negyedévre leszűrt értékesítési adatokra klaszterezést végrehajtva feltártuk az értékesítési rekordok egymáshoz viszonyuló kapcsolatát.

Ellentétben a felügyelt (*supervised learning*) gépi tanulással, ahol az egyes tanuló rekordok esetén valamilyen (nominális) célváltozó értéke már rendelkezésünkre áll, a klaszterezésnél, amely a nem-felügyelt tanulás (*unsupervised learning*) egyik változata, ilyen információval nem rendelkezünk. A legkézenfekvőbb megoldás az adatok természetes szerkezetének feltárása, pontosabban az egymáshoz hasonló rekordokat egyazon klaszterbe sorolni. A *K*-közép egy egyszerű partícionáló klaszterezési algoritmus, amely a megfigyeléseket *K* darab csoportba osztja, ahol *K* a felhasználó által megadott paraméter: a klaszterek száma. A *K*-közép minden egyes megfigyelést az ahhoz tartozó legközelebbi klaszterbe sorol, majd a klaszter középpontját frissíti a

Arres	Beszerzé.	Mennyiség	Nettó ela..	Vevő kód	Hónapok	Cikk kód	ClusterN..	Distance
223.97	12225.50	15.00	15585.00	V005700	2016.05.01.	24448	3	42.46
103.02	1487.95	3.00	1797.00	V007724	2016.04.01.	33010	5	34.79
0.00	0.00	0.00	0.00	V000243	2016.04.01.	22042	5	69.63
149.19	4098.10	10.00	5590.00	V005228	2016.06.01.	20132	5	80.96
199.15	1115.39	1.64	1442.00	V000058	2016.04.01.	44296	3	67.27
163.03	3571.20	6.20	4582.00	V007560	2016.04.01.	40299	5	94.81
114.26	1975.05	7.50	2832.00	V200030	2016.06.01.	23804	5	46.03
83.54	4309.30	20.00	5980.00	V001717	2016.06.01.	22580	5	15.31
119.21	6244.86	25.00	9225.00	V006566	2016.04.01.	22589	5	50.98
309.00	9810.00	9.00	12591.00	V007791	2016.05.01.	61637	3	43.99
379.52	3561.99	1.22	4025.00	V006556	2016.05.01.	42042	3	114.51
273.64	3395.36	1.00	3669.00	V006147	2016.05.01.	62364	3	8.63
128.86	992.28	1.48	1183.00	V003544	2016.06.01.	40397	5	60.64
111.20	3213.61	12.00	4548.00	V003322	2016.04.01.	20023	5	42.97
46.70	1390.00	10.00	1857.00	V200025	2016.04.01.	57368	5	22.93
172.01	18082.85	25.15	22409.00	v002710	2016.06.01.	44239	3	94.41
41.55	2034.51	10.00	2450.00	V006889	2016.06.01.	23777	5	28.09
209.70	30181.31	84.00	47796.00	V007429	2016.04.01.	10012	3	56.73

1. számú ábra: 2016/3. negyedévének klaszterezési eredményei
Forrás: saját szerkesztésű ábra, 2016.

klaszterhez rendelt pontok alapján mindaddig, amíg egyetlen pont sem vált klasztert és a középpontok ugyanazok nem maradnak. Ha a *K* értéke nem ismert előre, akkor különböző *K* paraméterű klaszterezések futtatására van szükség. A klaszterezést *K* [1, 10] értékekkel teszteltük, majd minden időszakra 5 klasztert hozunk létre (a többi paraméter esetén vagy túlságosan heterogén csoportok jöttek létre, vagy nehezen lett volna már értelmezhető a klaszterek magas száma), amelyek tartalma megmutatja, hogy mely eladási rekordok (cikk-vevő kombinációk) „hasonlítanak" egymásra a leginkább azok eladási árrései, valamint egyéb jellemzőik alapján (mint amilyen a beszerzési érték, eladott mennyiség, cikkcsoport stb.).

A klaszterezés paraméterezésekor megválaszthatjuk a klaszterek kialakításában részt vevő attribútumokat. A szerzők által kiválasztott attribútumok a következők voltak: *mennyiség, nettó sorösszeg, beszerzési ár és árrés*. Ezt követően a bemeneti adatokban található hiányzó értékeket figyelmen kívül hagytuk, az iterációk maximális számát pedig 100-ra állítottuk. A klaszterek kezdeti középpontjainak megválasztása nehéz kérdés, a tapasztalat azonban azt mutatja, hogy a legjobb eredményt úgy lehet elérni, ha a kiinduló középpontokat véletlenszerűen határozzuk meg. Ezentúl a Predictive Analytics különböző opciókat nyújt arra vonatkozóan, hogy a klaszterezési folyamat során hogyan számítsuk ki a középpontok új koordinátáit.

A klaszterek elkészítése előtt beállíthatjuk, hogy hány párhuzamos feldolgozási szálon történjen az adatok elemzése; amit a rendelkezésre álló erőforrások (CPU technológiája, teljesítménye alapján) függvényében növelhetünk. A számítási folyamat felgyorsítása céljából 2 szálon futtattuk az elemzéseket. Az utolsó, 2016.07.01 és 2016.10.01 közötti időszakra lefuttatott klaszterezési folyamat eredményét vizsgálva megállapítottuk, hogy a táblázatos megjelenítésben két új oszlop keletkezett: a klaszter sorszáma, mely megmutatja, hogy az adott rekord az 5 klaszter közül melyikbe került, valamint a távolság, amely az adott sor saját klaszterének középpontjától vett euklideszi távolságát mutatja (lásd **1. számú** ábra).

Az algoritmus befejezése után az 1. klaszter, amely középpontjának árrés koordinátája a legalacsonyabb, tartalmazza a legkevesebb tételt (lásd **2. számú** ábra). Ez arra utal, hogy kevés olyan eladás történt, ahol a beszerzési ár magasabb, mint az eladási ár. Például ez a közelgő lejárati idővel rendelkező termékek esetén lehetséges, hiszen előfordulhat, hogy a vállalkozás nem figyel oda megfelelően a pár hónapon belüli lejárati idővel rendelekző termékek megfelelő kezelésére. A legmagasabb árrés középpont-koordinátával (8978,69) rendelkező klaszterben (2.) viszonylag kevés termék található, ami arra utal, hogy a vállalat viszonylag kevés terméket értékesített magas nyereséggel.

```
Algorithm Summary
Summary:

Overview
-----------------------------------------------------
Model Building Date   :   10/31/16 1:38 PM

Independent Columns
    1. Arres : Double
Summary from In DB PAL Script:
-----------------------------------------------------

Number of clusters      :5

The size of each cluster:
Cluster1                                :21
Cluster2                                :30
Cluster3                                :49908
Cluster4                                :1697
Cluster5                                :130863

Sum of all clusters     :182519
Cluster Centers
                        Arres
1       :               -628.047619047618
2       :               8978.699538795805
3       :               265.71665871890696
4       :               1024.8347494683708
5       :               68.933937853437186
```

2. számú ábra: A klaszterezési algoritmus eredményének összefoglalója
Forrás: saját szerkesztésű ábra, 2016.

2. számú táblázat: A klaszterezési algoritmus eredményének összefoglalója

Klaszter száma	Klaszterben található rekordok mennyisége	Klaszter árrés szerinti középpontok	Besorolt kategória árrés alapján	Kumulált árrés	Besorolt kategória árbevétel alapján
1	21	-628	Nagyon rossz	-13188	Nagyon rossz
2	30	8 978	Nagyon jó	269340	Rossz
3	49 908	256	Semleges	13225620	Nagyon jó
4	1 697	1 024	Jó	1737728	Semleges
5	13 0863	68	Rossz	8898684	Jó

Forrás: saját szerkesztésű táblázat, 2016.

A klasztereket értékelési szempontból – nagyon jó, kevésbé jó, semleges, kevésbé rossz, nagyon rossz – úgy rendezhetjük, ha megvizsgáljuk az egyes klaszterek középpontjait, illetve a klaszterekbe sorolt termékek darabszámát, majd a darabszámot súlyoknak tekintve ezeket az értékpárokat összeszorozzuk, és ezen eredmények alapján rangsorolunk. Ezt az eljárást az utolsó negyedévre alkalmazva az alábbi eredményeket kaptuk *(*lásd **2. számú táblázat***)*.

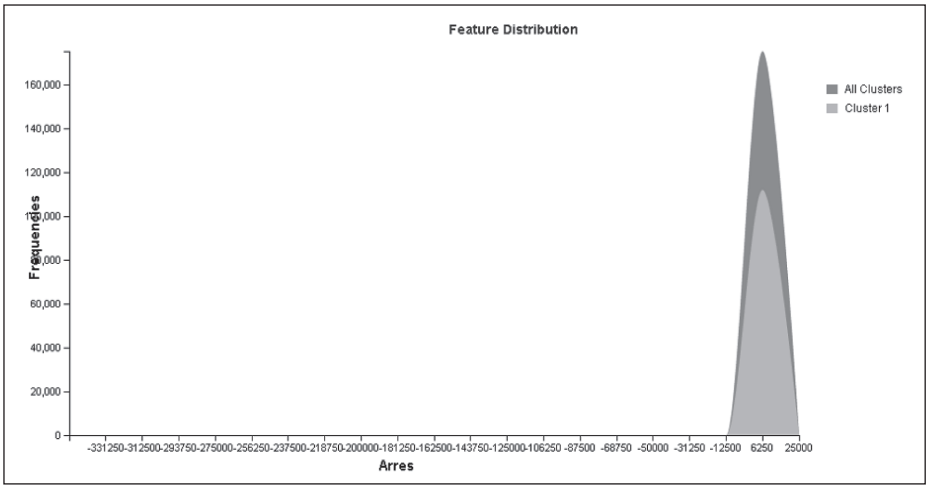
Az elemzések folyamán ezeket a középpontokat, tulajdonságokat és klaszter rendezéseket figyelembe véve célszerű következtetéseket levonni. Említést érdemel, hogy a klaszterek rangsorolására nincs túl sok kifinomult módszer a szakirodalomban (különös tekintettel ilyen jellegű, értékesítési adatokra). A klaszterekre vonatkozó tulajdonságokat különböző grafikai eszközök segítségével is szemléltetjük *(*lásd **3. számú ábra***)*.

Megállapíthatjuk, hogy a rekordok az egyes klaszterekben külön-külön is, illetve együttesen is az árrés attribútum szerint normális elosztást követnek, erős csúcsossággal és minimális ferdeséggel (amitől az eloszlások viszonylag szimmetrikusnak mondhatók).

A klaszterezés lefuttatását követően a helyettesítő termékekre vonatkozó vizsgálatokat az egyes cikkcsoportokra külön-külön is elvégeztünk, hiszen egy termék helyett nem feltétlenül javasolhatunk más cikkcsoportban lévő terméket. Adott időszakban két árrés alapján külön kategóriába került klaszterek cikkcsoportjait elemezve megállapítottuk, hogy a ,,kevésbé jó”› árrés középponttal rendelkező klaszterből kiragadott termékek helyett a vizsgált vállalat ajánlhatna olyan cikket, amely a ,,nagyon jó”› árrés középponttal rendelkező klaszterbe sorolható. Ez azt jelenti, hogy a vállalat a profitját a nagyobb árréssel rendelkező cikk alapján növelheti. Ilyen például, ha a ,,friss csirkemell” mellett ,,hűtött friss csirke mellfilét” is javasolunk a vásárlónak. Az előbbi termék alacsonyabb árréssel, míg az utóbbi magasabb árréssel rendelkezik. A vizsgált vállalat nagyon rossz kategóriába eső árréssel rendelkező termékeit kivonhatná a piacról, hiszen azok raktározása, selejtezése költséges lehet.

A klasztereket nemcsak egy negyedéven belül, hanem negyedévek között is összehasonlítottuk, melynek lényege a klaszterek idősorban történő változásának megvizsgálása. Az összes vizsgálatot elvégezve egy mátrixszerű kimutatásnak köszönhetően megállapíthatóvá vált a klaszterek szerkezetének negyedévről – negyedévre történő változása.

Továbbá megvizsgáltuk a két különböző, egymást követő időszak cikkeinek eladási mennyiségét, ami alapján arra a következtetésre jutottunk, hogy egy-egy cikk eladási tranzakciója nő, ennek köszönhetően a viszgált vállalat nyugodtan befektethet nagyobb mennyiségű készlet kedvező áron történő megvásárlásába. Ritkán fordul elő olyan eset, ahol a cikk értékesítése 0-hoz közelít, mégis ezen termékek kivezetése a piacról előrejutást jelenthet a vállalatnak. A klaszterek összehasonlítása segíthet a vállalat marketingjének megváltoztatásában, mely a magas árréssel rendelkező cikkekkel való kereskedést lendíthetné fel, vagy az egyes termékek készletezési stratégiáját javíthatná. Az adatbázisból meghatározható többek között az elmúlt hónapok alapján kiszámított 10 legjobb vevő listája is. Ezáltal a vevőket *gold* és *silver* kategóriába sorolhatjuk, így külön bónuszokkal, kedvezményekkel tarthatjuk meg hűséges vásárlóként.



3. számú ábra: Árrések eloszlása
Forrás: saját szerkesztésű ábra, 2016.

A klaszterezéseket az SAP Predictive Analytics operátorai mellett *R*-ben is lefuttattuk, amelyhez a következő kódrészletet készítettük (lásd **4. számú** ábra).

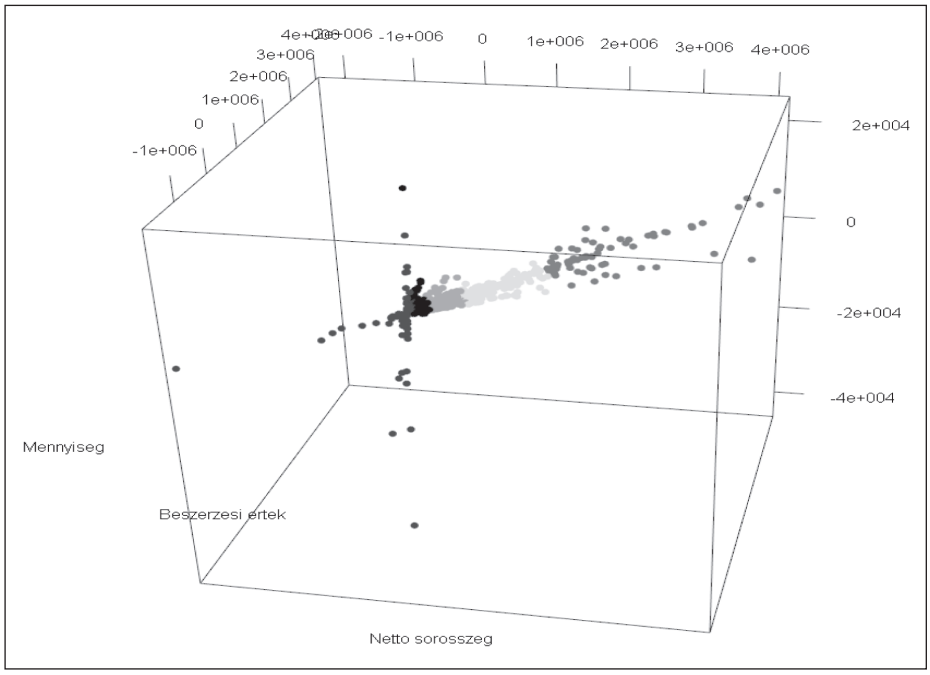
A fenti kódrészletben *data* jelöli az eredeti, negyedéves adatokat, melyeknek csak a folytonos attribútumait vettük figyelembe a klaszterezés során. Az iterációk számát, a Predictive Analytics paraméterezéséhez hasonlóan itt is 100-ra állítottuk. A klaszterezési eredmények a *results* változóba kerülnek, amelyekből az egyes rekordok klaszterazonosítóit cbind utasítással az eredeti adatokhoz csatoltuk, mint új, nominá-

```
# Klaszterezes futtatasa a folytonos attributumokra
results <- kmeans(data[,-c(1,2,3,4)], 5, iter.max = 100)

# Klasztereredmenyek rekordokhoz rendelese:
tableWithCID <- cbind(data, results$cluster)

# Az eredmenyek vizualizalasa:
library(rgl)
plot3d(tableWithCID$NettoSor,
       tableWithCID$BeszErtek,
       tableWithCID$Mennyiseg,
       col = tableWithCID$`results$cluster`,
       size = 8,
       xlab = "Netto sorosszeg",
       ylab = "Beszerzesi ertek",
       zlab = "Mennyiseg")
```

4. számú ábra: R programnyelv K-közép klaszterezési kódja
Forrás: saját szerkesztésű ábra, 2016.



5. számú ábra: Klaszterezési eredmények 3D-s grafikonon
Forrás: saját szerkesztésű ábra, 2016.

lis attribútumot. Végül az eredményeket grafikus formában a *plot3d(...)* utasítás jeleníti meg (lásd **5. számú** á**bra**).

Az **5. számú** ábra a 2016/3-as negyedév értékesítési rekordjainak *K*-közép algoritmussal kapott 5 klaszterét szemlélteti. Látható, hogy az eljárás viszonylag sikeresen választotta szét a negyedév adatait: a kiugró, jobban szóródó kék, valamint piros színnel jelölt halmazok elemei 1-1 klaszterbe kerültek, illetve a tér közepén elhelyezkedő fekete, zöld és türkiz színű csoportok is elkülönülnek egymástól. A kék színű klaszterbe eső rekordokhoz tartozó cikkeket javaslatunk szerint indokolt felülvizsgálni, a vállalat kínálatából azokat nagy szóródásuk, alacsony nettó sorösszegük miatt elhagyni. A fekete, zöld és türkiz színekkel jelölt csoportokhoz tartozó cikkekre a vállalatnak célszerű nagyobb figyelmet fordítania, ugyanis ezek viszonylag stabil kereskedési, bevételi teljesítményt nyújtanak, a jövőben pedig könnyen képviselhetik a piros színnel jelölt klaszter elemeit, amelyek a vállalat legeredményesebb értékesítési adatait reprezentálják.

Kutatásunkban többnyire a *K*-közép algoritmushoz ragaszkodtunk és nem fordítottunk különösebb figyelmet egyéb klaszterezési lehetőségekre. Egyéb módszert is kipróbáltunk, melyek közül talán a legfontosabb a DBSCAN eljárás. Ez alapvetően egy sűrűség alapú klaszterezési módszer, de a kialakított klaszterek száma előre nem meghatározott, így a sűrűség definiálásához szükséges két paraméter, a sugár *(eps)* és a közelségi pontok *(minPts)* variálása problémát okozott a kapott klaszterek mennyiségét illetően (bizonyos paraméterezési kombinációkkal 50-100 körüli klaszterszámot is kaptunk, aminek az értelmezése meglehetősen nehéz), de az is gondot okozott, hogy minden értékesítési rekordot sikeresen besoroljunk valamely klaszterbe (sok paraméterezés esetén az értékesítési rekordoknak egy jelentős részét a módszer kiugró adatoknak találta). Végezetül az eljárás lényegesen erőforrásigényesebb volt, mint a *K*-közép eljárás (*eps*=1000 és *minPts* =10 esetén az eljárás megközelítőleg 9 gigabájt memóriát fogyasztott).

Az előrejelzések eredményei

Az előrejelzések készítése során különböző regressziószámítási eljárásokkal modelleztük az egyes klaszterek középponti koordinátáinak időbeni alakulását, valamint az idősor-analízisbe a klaszterek többi jellemzőjét is bevontuk.

A statisztikailag megalapozott következtetések levonásához „sok"' adatra van szükség, így pl. ha a klaszterek idő függvényében vizsgált középpontjait szeretnénk előre vetíteni, akkor szükséges lehet legalább 3, de inkább 4 év negyedéves klaszterezés eredményeire. Az előző évek historikus adatait figyelembe véve elmondhatjuk, hogy a vizsgált vállalat felkészülhet a decemberi csúcsforgalmára, valamint az azt követő januári bevétel csökkenésére. Így olyan előrejelzéseket készítettünk, amely a jövőre nézve akár a szezonális termékek felkészülésére is javaslatot ad a vizsgált vállalat számára. A rendelkezésre álló 15 negyedév adatainak *K*-közép klaszterezéssel kapott eredményeire, pontosabban a klaszterek középpontjaira lineáris regresszió segítségével előrejelzéseket készítettünk. Ezen eljárást formálisan az alábbiak szerint fejezhetjük ki. Jelölje C_i egyes negyedévekre kapott 5 klaszter középpontjainak halmazát, ahol az i {1, 2, ...15}, azaz a 15 negyedév külön-külön vett klaszterezési részeredményét (tehát jelenleg csak a prototípusokkal foglalkozunk, azaz a klaszterek középpontjaival, figyelmen kívül hagyva az egyes rekordok új klasztercímkéit).

Amennyiben a klaszterek kialakításánál a felhasznált dimenziók az árrés, mennyiség, nettó sorösszeg és beszerzési ár, úgy a klaszterek középpontjait is ezek a folytonos attribútumok fogják képviselni. Ekkor, egyetlen negyedév klaszterezési eredményének ezen részét (középpontok) az alábbi mátrixszal jelölhetjük:

$$C_i = \begin{pmatrix} c_{1,1} & \cdots & c_{1,4} \\ \vdots & \ddots & \vdots \\ c_{5,1} & \cdots & c_{5,4} \end{pmatrix}$$

ahol $c_{i,j}$ az *i* klaszter középpontjának *j*. koordinátája. Mivel 15 negyedév elemzési eredményeivel rendelkezünk, 15 darab ilyen mátrix fogja képezni a regressziós elemzés alapját. A kérdés már csak az, hogy mely klaszterek mely középponti koordinátáinak alakulását modellezzük, valamint az, hogy mely változókat vonjuk be a modellbe, mint magyarázó változókat. Magyarázó változóként az időt építettük be a modellbe. Ezen változó értékei a természetes számok az (1, 15) intervallumban.

A modell bonyolítása során a következő lehetőségeink vannak: elsőként bevehetjük a C_i mátrixok (i 1, . . .,15) megfelelő sorainak kimaradt oszlopait, azaz a vizsgálni kívánt klaszter középpontjának többi koordinátáját. Végezetül a modell további 16 magyarázó változó bevonásával építhető tovább, ugyanis ha magunk elé képzeljük a fenti C_i mátrixokat, akkor láthatjuk, hogy azoknak 5 sora és 4 oszlopa van (az elkészített klaszterek száma és középpontok koordinátáinak mennyisége miatt). Mivel egy adott klaszter középpontjának egy kiválasztott koordinátája már célváltozóként szerepel a modellben és az adott klaszter középpontjának maradék 3 koordinátáját már így is bevontuk, a mátrixnak 4 sora maradt ki (azok minden oszlopával), így 16 további független változó áll rendelkezésünkre.

A **3-6. számú táblázatokban** szemléltetjük a függő és független változók kiválasztását.

A egyes sorok azt jelölik, hogy mely klaszter középpontjáról van szó, az oszlopok pedig a középpontok egyes koordinátáit reprezentálják: M, N, B és Á rendre a mennyiség, nettó sorösszeg, beszerzési ár és árrés attribútumok rövidítései. A **3-6. számú táblázatok** mindegyike esetén a piros színű cella jelöli a célváltozót; a **4. számú táblázat** azt az esetet hivatott szemléltetni, amikor csak a vizsgált klaszter koordinátáival dolgozunk, míg a **5. számú táblázat** esetén csak egy koordinátával foglalkozunk, azonban az összes klaszter adatát figyelembe vesszük (értelemszerűen, a **6. számú táblázatnál** mindezen magyarázó változók a modell részét képezik).

Nagyon fontos kérdés továbbá, hogy hogyan állapítsuk meg az egyes negyedévek között az összetartozó klaszterpárokat: mivel alapvetően itt egy nem-felügyelt tanulási feladatról van szó, az 5-5 klaszter negyedévről negyedévre történő összeláncolása (így képezve a klaszterek középpontjaiból idősorokat) nem triviális. Igyekeztünk kifinomult módszert választani, nevezetesen a Magyar-módszert, amelyet az operációkutatás területén széles körben alkalmaznak, például hozzárendelési feladatok megoldására. Ehhez nem szükséges más, csupán az

3. számú táblázat:Kétváltozós regresszió

	M	N	B	Á
1.				
2.				
3.				
4.				
5.				

5. számú táblázat:Többváltozós reg regresszió (2)

	M	N	B	Á
1.				
2.				
3.				
4.				
5.				

4. számú táblázat: Többváltozós regresszió (1)

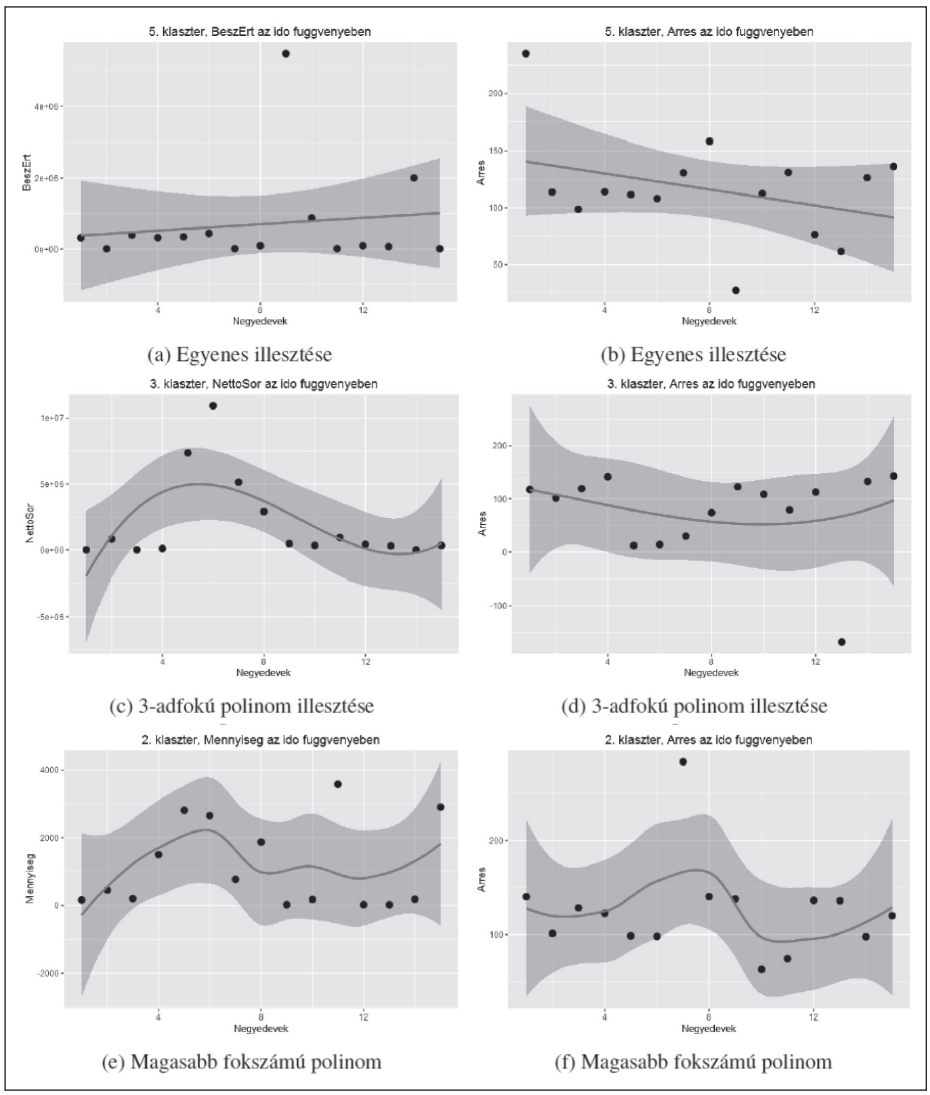
	M	N	B	Á
1.				
2.				
3.				
4.				
5.				

6. számú táblázat: Többváltozós r regresszió (3)

	M		N	B	Á
1.					
2.					
3.					
4.					
5.					

3-6. számú táblázatok

Forrás: saját szerkesztésű táblázatok



6. számú ábra: A regressziós modell eredményei
Forrás: saját szerkesztésű ábra

aktuális két negyedév klaszterközéppontjainak egymástól vett távolságait kiszámítani, ami azt jelenti, hogy az 5-5 klaszter miatt egy mátrixot fogunk kapni. Végül az így kapott mátrixra alkalmaztuk a Magyar-módszert, amely kimenetként megadta, hogy az előző és a következő negyedév klasztereit hogyan párosítsuk. Formálisan: Magyar-módszer , ahol valamint páros az egyes klaszterhalmazok elemeinek összepárosítása (pl. *i=3, j=4* esetén a következő negyedéves elemzésekből a 4. klasztert feleltetjük meg az előző negyedév 3. klaszterének). Ezt az eljárást negyedévről negyedévre alkalmazva a problémának egy optimális megoldását találtuk meg, ami a leginkább összeillő elemekből felépülő idősorokat eredményezte számunkra.

Az alábbiakban néhány példával szemléltetjük azokat az eredményeket, amelyeket a fenti módon kialakított idősorokra regresszióelemzéssel kaptuk. Mivel a magasabb (több, mint három) dimenziójú tereket vizualizálni nem igazán lehetséges, ezért itt kizárólag 2-változós regressziós eredményeket ismertetünk: olyan eseteket, amelyekben az egyetlen modellbeli független változó az idő, a célváltozó pedig valamely klaszternek valamely középponti koordinátája. A klaszter-középpontok erős ingadozása miatt a lineáris regresszió csak néhány esetben tudta jól megfogni az idő és az adott koordináta kapcsolatát, így az esetek többségében egy 3-adfokú polinommal realisztikusabban modelleztük az adott idősort. Fentebb említettük, az R beépített eszközeinek köszönhetően lehetőségünk van a kész függvényekre bízni a legjobban illeszkedő, de nem túlillesztett polinom fokszámának meghatározását. Ezek az illesztések már kimondottan alacsony hiba-négyzetösszeggel teljesítettek, amit az alábbi **6. számú** ábrá**n** megfelelően szemléltetünk. Figyelmet fordítottunk még a regressziós modellek jóságának vizsgálatára is. Röviden elmondhatjuk, hogy a regressziós modelleken végzett, az egyes együtthatók létjogosultságára irányuló tesztek és a modelleket egészében vizsgáló, globális F-próba is elfogadható eredményeket mutatott. Az esetek többségében mind az idő, mind az egyéb független változók szignifikáns hányadát magyarázták a kiválasztott célváltozó varianciájának. Az alábbi regressziós eredményeket célszerű óvatosan kezelni: a hosszú távú extrapoláció nem ajánlott, ugyanis az illesztett görbékkel való prediktálás több negyedévre már pontatlan lesz (a modellezéssel kapott függvények csak a megfigyelések intervallumára, valamint annak szomszédságára adnak viszonylag pontos becsléseket).

A regresszió számítás eredményeit a **6. számú** ábrákon szemléltetjük.

A fenti 6 regressziós modell eredménye látható, amelyek közül az első kettő egy-egy elsőfokú polinom illesztéséről szól. Az egyeneseknek néhány adatponttal vett négyzetes hibája viszonylag magas, ugyanis egy ilyen egyszerű modellel alig lehet megteremteni a kapcsolatot két változó között. Értelmezésük a következő: **6.a. számú** ábrán az 5. klaszter beszerzési érték koordinátájának alakulását látjuk az idő függvényében, ami a modell szerint növekvő trendet követ, ezzel szemben viszont az árrés csökken (**6.b. számú** ábra), mint ahogyan az várható volt. A vállalatnak célszerű ezen klaszter eladási rekordokhoz tartozó termékeit jobb árazási stratégiával ellátni, ugyanis az előrejelzés szerint (a lineáris modellek az esetek többségében jobban általánosítanak, így becsléseik megbízhatóbbak) néhány negyedéven belül a terméken realizált nyereség minimalizálódni fog. A **6.c. és 6.d. számú** ábrán már harmadfokú polinomok illesztésének eredménye látható, ami a 3. klaszter negyedéves eredményeinek nettó sorösszeg, valamint árrés koordinátáinak becslését hivatottak elvégezni. A klaszterbe eső értékesítési adatok a vizsgált időszak közepén növekvő tendenciát mutattak a nettó eladási árat tekintve, jóllehet, a vállalat rajtuk realizált nyeresége az árrés szerint a közbülső negyedévekben csökkent.

Ezt részben a beszerzési érték erősebb növekedésével, részben a klasztert váltó termékek fluktuációjával magyarázhatjuk, ezért a vállalatnak az ide sorolt termékek közül a kockázatosabbakat indokolt elvetnie, a kevésbé ingadozókat pedig stabilizálni (ezt a megkülönböztetést pl. az adott termékek cikkcsoportjába tartozó termékek megvizsgálásával is megalapozhatjuk).

Összegzés, következtetések és javasalatok

Az SAP HANA-nak köszönhetően egy újszerű technológián alapuló adathalmazt hoztunk létre (nevezetesen egy OLAP kockát), ezt követően az adatok tárolási módja miatt elemeztük az értékesítési rekordokat. A klaszterezés lefuttatását követően helyettesítő termékeket fedeztünk fel a vállalat egyes cikkcsoportjain belül. Ha a vállalat az ilyen jellegű termékek közötti kapcsolatokra jobban odafigyel és marketing stratégiájában fokozottan koncentrál a magasabb árréssel rendelkező termékekre, akkor rövid időn belül nagyobb jövedelemre tehet szert. A helyettesítő termékeken kívül szezonális termékek is megtalálhatók a klaszterekben, melyeket az egyes időszakokban különböző eladási mennyiség jellemzett. A vállalat az ilyen jellegű ingadozásokra úgy készülhet fel, hogy ezekből a cikkekből az növekvő periódus előtt nagy mennyiséget, akciós áron szerez be. Továbbá a klasztereket külön kategóriába sorolva találtunk „nagyon rossz”› árréssel rendelkező csoportokat is. Az ilyen klaszterben található cikkek vagy közelgő lejárati idővel rendelkező termékek az úgynevezett „befagyott”› készletek. E cikkek koordinálására a vállalatnak célszerű nagyobb figyelmet fordítani, vagyis az SAP Business One rendszer képességeivel „figyelmeztetési eljárásokat” indokolt használnia.

Általában de elsősorban a top 10 vevőre vonatkozóan olyan marketing stratégiát alakíthat ki a vizsgált vállalat, ami megtartja, sőt fellendíti a vevők vásárlási szokásait. Ezt úgy érheti el, hogy például a hónap első két hetében legtöbbet vásárolt vevőknek a számla nettó végösszegéből egy előre meghatározott kedvezményt biztosít, ezáltal a vevők egyre többet vásárolnak.

A klaszterezést az R programnyelvvel is elvégeztük, melynek köszönhetően a klaszterezési eredményeket, vagyis az egyes értékesítési rekordokat a hozzárendelt klasztercímke alapján színezve, 3D-s formában is vizualizáltuk. Ezen grafikus kimenetekre tekintve megállapítottuk, hogy vannak olyan termékek, amelyeket célszerű kivonni a piacról, míg másokat indokolt fejleszteni, vagy készletezésének minimum szintjét megemelni.

A regresszióanalízis során az egyes klaszterek mozgását jeleztük előre, elsősorban az idő függvényében, mind egyszerűbb regressziós technikákkal, mind bonyolultabb modellekkel. Megállapítottuk, hogy a bemutatott technológiákban és az adatbányászati algoritmusokban rejlő lehetőségek tárháza végtelen. Habár az elmúlt években mind az adattárolási technológiák, mind az adatbányászat tudománya hatalmas fejlődésen ment keresztül, még közel sem tartunk ott, hogy ezeket a módszereket teljes biztonsággal és pontossággal alkalmazzuk kis- és középvállalati környezetben.

Megállapítottuk, hogy a vállalat a javaslatok elfogadása mellett folytathatja meglévő, sikeres versenystratégiáját a technológia folyamatos fejlesztésével párhuzamosan, különös tekintettel azon belül a raktározási, logisztikai, szállítmányozási erőforrásaira.

Felhasznált irodalom

AHSAN, A. (2009): Analysis of mealybug incidence on the cotton crop using ADSS-OLAP Online Analytical Processing tool. Computers and Electronics in Agriculture, Volume 69, Number 1, p. 59-72.

GARCIA-ALVARADO, C. – ORDONEZ, C. (2015): Clustering binary cube dimensions to compute relaxed GROUP BY aggregations. Information Systems 53, p. 41-59.

HASSO, P. (2012): A Common Database Approach for OLTP and OLAP Using an In-Memory

Column Database. Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, p. 1-2.

HELEN, T. – ANINDYA, D. (2001): A Conceptual Model and Algebra for On-Line Analytical

Processing in Decision Support Databases. Information Systems Research, Volume 12, Number 1, p. 83-102.

MARK, W. (2013): Software Development on the SAP HANA Platform. Pack Publishing Ltd., Birmingham

MICHELANGELO, C. – ALFREDO, C. – DONATO, M. (2015): Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering. Journal of Intelligent Information Systems, Volume 44, Number 3, p. 309-333.

NIELS, N. – STEFAN, M. – SJOERD, M. – MARTIN, K. (2012): MonetDB: Two Decades of Research in Column-oriented Database Architectures. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Volume 35, Number 1, p. 40-45.

OKSANA, G. – JEROME, D. – JEAN-HUHUES, C. – IRYNA, Z. (2010): Business intelligence

for small and middle-sized entreprises. ACM SIGMOD Record, Volume 39, Number 2, p. 39-50.

OMAR, B. – MOHAMED, H. – ABDESSADEK, T. – TARIK, A. (2015): Multi-criteria Decisional Approach of the OLAP Analysis by Fuzzy Logic: Green Logistics as a Case Study.

Arabian Journal for Science and Engineering, Volume 40, Number 8, p. 2345-2359.

PENNY, S. – ROB, F. – BJARNE, B. (2015): SAP HANA An Introduction. Pack Publishing Ltd., Birmingham

SAP SE (2014): SAP HANA Developer Guide. SAP affiliate company, Waldorf, Document Version: 1.1

WOLFGANG, N. (2009): SAP Business One Implementation. Pack Publishing Ltd., Birmingham