5. Adatbányászok mintafeladata: prediktív modellezési probléma

„Vitathatatlanul előnyhöz jut az,
aki pontosabban ismeri a majd bekövetkező esemény
valamilyen aspektusát.”

5.1. Jövőbe látni

Az adatbányászati tevékenység – mint minden modellezés – célját tekintve lényegében két alapvető csoportra bontható:

- konstruktív modellezés (adatbányászat), illetve
- prediktív modellezés (adatbányászat). |103|

Az előbbi egy létező rendszer (gazdasági, műszaki, szociológiai stb.) belső működési mechanizmusainak feltárását szolgálja, míg a második egy létező vagy hipotetikus rendszer jövőbeli működésének *előrejelzését* célozza. Ez utóbbit *prediktív adatbányászatnak* nevezzük.

Az adatbányászat sokak szerint legizgalmasabb területei azok, amelyek segítségével a jövő eseményeivel kapcsolatos megállapítások tehetők. Ennek a kijelentésnek az oka egyszerű: az emberi kíváncsiság mindenre ráirányul, ami valamilyen talányt rejt magában. A kíváncsiság szülte számtalan kérdés között pedig van egy, ami így hangzik: „Mi fog történni?". És lássuk be, ez szinte kivétel nélkül mindenkit foglalkoztat valamilyen formában.

Ez a kérdés olyannyira az életünk része, hogy legtöbbször észre sem vesszük, hogy a jövőbeli események alakulásán spekulálunk. Közlekedünk és közben, részben tudattalanul, folyamatosan a környezetünkre figyelünk, apró előrejelzéseket teszünk a közlekedésben részt vevő egyéb szereplők mozgására, állapotára vonatkozóan. Az otthonról indulás előtt kinézünk az ablakon, és döntést hozunk: betesszük a táskánkba az esernyőnket vagy sem. Vitatkozunk, és megpróbáljuk előre kitalálni vitapartnerünk reakcióit a saját mondandónk, valamint a vitapartnerünk értelmi képességei, habitusa, lelki állapota alapján. Tekegolyót gurítunk, és gurulás közben folyamatosan elemezzük a golyó pályáját, és már a bábukhoz való megérkezése előtt gyanítjuk, hogy sikerült-e jól gurítanunk vagy sem.

A jövő várható történéseinek fürkészése sok esetben viszont nagyon is tudatos tevékenység, kiváltképp versenyhelyzetben. Vitathatatlanul előnyhöz jut az, aki *pontosabban* ismeri a majd bekövetkező esemény valamilyen aspektusát.

A jövő pontosabb előrejelzésének képessége azonban csak befektetések árán szerezhető meg. Ez a befektetés mindig valamilyen tanulást és elemzői munkát jelent, aminek az eredménye egy „képlet". A kis gyermek, aki éppen csak ismerkedik a világgal, gyűjti az impulzusokat, és igyekszik összerakni saját képleteit, amelyek persze élete első szakaszában egyszerűek lesznek („Ha sírok, akkor megjelenik anya, és ölbe vesz" vagy „Ha a tárgy gömbölyű és meglököm, akkor elgurul."). Ugyanezzel a módszerrel igyekszik a pókerjátékos is összerakni a képleteit („Ha sokat matat a kezével, akkor valószínűleg jó lapjai vannak."), de analógiát mutat ezzel az esernyővel vagy anélkül való otthonról indulás szituációja is. Honnan sejtjük, hogy esni fog az eső? Rengetegszer láttunk már borult égboltot, és megtanultuk (a tudattalanul bennünk működő elemzéseink során felismertük), hogy a borult égbolt és az eleredő eső ok-okozati viszonyban állnak egymással.

A képleteink az életünk folyamán változnak, két okból kifolyólag. Egyrészt finomodnak, a folyamatosan érkező újabb és újabb tapasztalatok hatására. Az újonnan tanultakat felhasználjuk, ezáltal csiszoljuk a jövőt előrevetítő képességünket. A kezdő autóvezető idővel megtanulja, hogy hogyan bánjon az autójával, és tudni fogja, hogy milyen mozdulatokra az hogyan fog reagálni. A felcseperedő gyermek megszokja,
|104| hogy nem minden szőrös állat olyan barátságos, mint az, amelyikkel otthon játszott a kertben. Sőt felnőttkorára az addig megismert állati viselkedési mintákból egész képletrendszere lesz, és tudni fogja, hogy melyik állattól milyen viselkedésre számíthat.

Másrészt a környezet időközben megváltozhat, és ha ehhez nem alkalmazkodunk, a régen felállított szabályok érvényüket vesztik. A mai fővárosban hullámzó forgalom minden bizonnyal szokatlan és kiszámíthatatlan lenne egy évtizedekkel korábbról időutazással idekerülő sofőr számára, aki az akkori közlekedési minták alapján bontakoztatta ki a közlekedési helyzeteket előrejelző képességét.

5.1.1. Az előrejelző modellezés

Az adatbányászat prediktív (előrejelző) modellezéssel foglalkozó területe az emberi tanulást utánzó módszerekkel – de az emberi tanulásnál sokkal gyorsabban – próbálja előállítani azokat a képleteket, amelyek egy-egy jól definiált környezetben képesek lesznek a „jövőbe látni", illetve képesek lesznek *pontosabb* előrejelzést adni (például egy emberi szakértőnél). A gyorsaság abból adódik, hogy a tanuláshoz alapanyagul szolgáló minták, megfigyelések adatbázisokban már rendelkezésre állnak, míg a tanulást végző algoritmusok a számítógépek másodpercenkénti sok millió műveletes sebességét használhatják.

Az előrejelzések pontosságát illetően nem szabad, hogy illúzióink legyenek. Előrejelzéseink csak annyira lehetnek megbízhatók, amennyire megfigyeléseink komplexitása kiterjedt. Ha csak abból tudunk következtetni arra, hogy lesz-e eső vagy sem, hogy van-e felhő az égen (és ezen kívül semmilyen más szempontot nem veszünk figyelembe), nyilván kevesebbszer találjuk el az igazságot, mint ha módunkban áll a hőmérsékletet és a légnyomást is figyelembe venni. Mindehhez ráadásul elengedhetetlen, hogy birtokában legyünk egy olyan tudásnak, egy olyan képletnek, amely maximálisan figyelembe tudja venni mindhárom szempontot. Az adatbányászat éppen ezt kínálja, azaz a rendelkezésre álló megfigyelések kiterjedtsége által szabott határokon belül a lehető legjobb képletet, a legjobb modellt építi fel, amellyel az előrejelzések megtehetők.

5.1.2. Példa az előrejelző modellezésre

Nézzünk egy konkrét példát! Ha egy hitelek kihelyezésével foglalkozó vállalat tudni szeretné, hogy a hozzá hitelért folyamodók mennyire lesznek megbízható adósok, akkor a következőt teheti az adatbányászat, a prediktív modellezés segítségével. Összegyűjti a korábban hitelért folyamodók elérhető adatait, azokét, akik pontosan fizetik a hiteltörlesztést, és azokét is, akiknek késedelmük van. Modellt épít, amely a rendelkezésre álló szempontok alapján (az adós életkora, családi állapota, iskolai végzettsége, fizetése, a hitel nagysága stb.) még a hitelfelvétel előtt minősíteni képes a hiteligénylőt. A modell mint az előrejelzés képlete magában fogja hordozni azt a tudást, hogy mely szempontokat milyen súllyal kell figyelembe venni az adós előminősítéséhez. A befektetett munka abban térül meg, hogy a modellt felhasználva pontosabb előrejelzést tehet a vállalat, mint ha csupán a saját, humán szakértőit kérte volna fel az ügyfelek előzetes megítélésére. Ezáltal kevesebb visszafizetetlen hitele lesz, amiből bőven megtérül az előrejelzésre fordított energia. |105|

Tisztában kell azonban lenni azzal a korláttal, hogy a hitel-visszafizetési hajlandóság gyakran nem az életkoron, családi állapoton, iskolai végzettségen, fizetésen vagy a hitel nagyságán múlik. Sokszor váratlan élethelyzetek, azaz olyan szempontok okozzák egy hitel bedőlését, amelynek figyelésére a hitelező vállalat, így a prediktív modell sem lehet felkészülve. A modelltől tehát nem várhatunk „tökéletes” előrejelzést, csak „jobbat”, mint amit anélkül tudtunk volna tenni.

A fentiekben vázolt hitelkérelmi probléma könnyen általánosítható a következő feladatra: adott egy ügyfél és adott egy eldöntendő kérdés. Tudni szeretnénk, hogy az adott ügyfélnél az adott kérdésre várható válasz igen vagy nem lesz. Összpontosítva az előbbi konkrét problémára: az ügyfél vissza fogja fizetni a hitelt vagy sem?

Mivel ez a problématípus gyakran kerül az üzlet fókuszába, ezért az üzleti haszon termelésének igájába fogott adatbányászat természetesen szolgál megoldással a feladatra. Sőt olyannyira fontos területe ez az adatbányászatnak, hogy a szerzők a következőkben bemutatott, a probléma megoldására szolgáló modellezési módszertant tekintik az adatbányászat alap- és mintafeladatának.

5.2. A prediktív modellezés folyamata

Az előző, adatbányászati módszertanokat taglaló 4. fejezetben bemutattunk egy CRISP-DM nevű szabványt, amely leírja az adatbányászati folyamatok felépítését. Látható volt, hogy a folyamat az alábbi lépésekből áll:

- üzleti környezet megismerése,
- adatok megismerése,
- adatok előkészítése,