SZAKDOLGOZAT

Tóth Zoltán

BUDAPESTI CORVINUS EGYETEM
ADATELEMZÉS ÉS INFORMATIKA INTÉZET
INFORMÁCIÓRENDSZEREK TANSZÉK

EGY BANKI HITELBÍRÁLATI RENDSZER LEHETSÉGES MODELLJE

TÓTH ZOLTÁN

ADATELEMZŐ SZAKIRÁNYÚ TOVÁBBKÉPZÉS

2022

KONZULENS: DR. VARGA KRISZTIÁN EGYETEMI DOCENS

TARTALOMJEGYZÉK

Tartalomjegyzék 3

1 Bevezetés 5

2 Kockázatkezelés egy hitelintézetben 7

2.1 Minősítő rendszerek helye a banki kockázatmenedzsmentben 7

2.2 Szokásos fejlesztési folyamat 12

2.2.1 Adatgyűjtés, mintavétel, definíciók 12

2.2.2 Adatelőkészítés 14

2.2.3 Feltáró adatelemzés 14

2.2.4 A modell összeállítása 15

2.2.5 Modell kiértékelése 16

2.2.6 Implementáció 16

3 A feladat meghatározása 17

4 Leggyakoribb osztályozási technikák 18

4.1 Lineáris regresszió 19

4.2 Logisztikus regresszió 20

4.3 Döntési fák 22

4.4 Diszkriminancia analízis 26

4.5 KNN 27

4.6 Neurális hálók 28

4.7 SVM 29

4.8 Hibrid módszerek 30

5 Adathiányok kezelése 32

6 Változószelekciós eljárások 35

7 Kiértékelési eljárások 39

8 Problémák a scoring modell építésénél 42

9 Egy minta-modell kialakítása R-ben 44

9.1 Az adatbázis bemutatása 44

9.2 Feltáró adatelemzés, adatok előkészítése 46

9.3 Belső összefüggések, információs tartalom 50

9.4 Modellek felépítése és kiértékelése 52

9.4.1 Döntési fa 52

9.4.2 Logisztikus regresszió 55

9.4.3 Modellek összehasonlítása 57

10 Összefoglalás 58

11 Felhasznált irodalom 60

1 BEVEZETÉS

A pénzügyek, azon belül is a hitelintézetek, a hitelezés világa az, ahol az adatelemzési módszerek már évtizedekkel ezelőtt megjelentek és gyorsan elterjedtek. Napjainkban is ez az egyik üzleti szféra, ahol széles körben, a napi gyakorlatba építve használnak adatbányászati módszereket.

Egy professzionális hitelső számára kulcsfontosságú, az eredményességet, sikert alapvetően meghatározó kérdés, hogy helyes döntést hozzon, amikor egy-egy ügyfél hitelképességéről dönt, illetve amikor folyamatosan vizsgálja, elemzi portfolióját, próbálja megbecsülni kockázati költségét, árazza az egyes termékeket, sőt az egyes ügyleteket. Mindezt úgy teszi, hogy figyelembe veszi startégiáját és az üzleti-gazdasági környezetét és annak előre látható változásait is. Mindemellett meg kell felelnie számtalan jogszabálynak és más előírásnak, alá kell támasztania tőkeallokációját. Ma már ez elképzelhetetlen adatelemzési modellek nélkül.

Jómagam több mint 25 évet eltöltöttem a finanszírozásban, magam is láttam ezt a hátteret és a folyamatos fejlődést, így magától értetődő volt, hogy amikor szakdolgozati témát kellett választanom, hogy egy ehhez kapcsolót jelöljek meg.

Természetesen egy szakdolgozat nem vállalkozhat a terület átfogó bemutatására. Célom az, hogy betekintést adjak, hol helyezkedik el egy scroring rendszer a banki kockázatkelés nagy rendszerében, milyen modellek honosodtak meg a gyakorlatban, azoknak mik a legfontosabb tanulságaik. Próbálom bemutat egy scorecard fejlesztésének lényegesebb lépéseit, és mindezt egy saját, egyszerű modell kialakításával is demonstrálni kívánom. A modell kialakításakor egy lakossági, személyi kölcsönt elbíráló „rendszert” készítek, de annyiban is egyszerűsítve az eljárást, hogy az csak a döntést támogatja meg (igen/nem),

tehát egy binominális klasszifikációt ad, és eltekintek a kockázati paraméterek becslésétől (PD, LDG). Azt feltételezem, hogy ez az egyszerű demonstráció is bemutatja az ilyen adatbányászati módszerek hasznosságát, de feltételezem azt is, hogy a nehézségeket, a fejlesztés során felmerülhető buktatók egy jó részét is.

2 KOCKÁZATKEZELÉS EGY HITELINTÉZETBEN

2.1 MINŐSÍTŐ RENDSZEREK HELYE A BANKI KOCKÁZATMENEDZSMENTBEN

Egy hitelintézet működésének lényege, hogy prémium ellenében kockázatokat vállal, tehát arra törekszik, hogy a tevékenységből fakadó veszteségforrásokat és jövedelmeket egyensúlyban tartsa. A kockázatokat feltáró, mérő, kezelő rendszer a kockázatkezelés, a bankmenedzsment kulcsterülete. Legfontosabb elemei (KOVÁCS - MARSI 61. o. alapján):

- A kockázatok azonosítása: a potenciális veszteségforrások és az üzletmenetet veszélyeztető lehetséges események feltárása
- A kockázatok mérése: ez magában foglalja ez erre szolgáló mérési módszerek és rendszerek kialakítását és a ezek működtetését; ide tartoznak a különböző minősítő rendszerek, tehát a scoring is, valamint a limitek, eljárások.
- A kidolgozott módszertanok érvényesítése a folyamatokban (pl. hitelelbírás rendje, de gyakorlatilag minden banki folyamatban)
- Kockázati monitoring: a bank kockázati pozíciója folyamatos nyomon követése, tartalékok, tőkekövetelmények számítása, stb.
- Kockázatcsökkentési eljárások: folyamatba épített ellenőrzések, limitek, stb.

Ezt a nagyon bonyolult és kifinomult tevékenységet és csak annyiban szeretném bemutatni, ami szükséges ahhoz, hogy a hitelbírálati módszerek, a scoring jelentősét megértsük és elhelyezzük ebben a rendszerben.

A három legfontosabb kockázattípus, ami egy hitelintézetnél / pénzügyi vállalkozásnál fellép. az a

- hitelezési (az adósok egyedi fizetési kockázata, ágazati kockázatok)

- piaci (árfolyamok, kamatok)
- működési kockázat (csalás, IT biztonság, folyamatok).

Emellett még a likviditási és az egyéb kockázatok csoportját (reputációs, stratégiai, stb.) szokták még kiemelni.

Mint köztudott, a banki működés az egyik legjobban szabályozott tevékenység; a különböző szabályok egyik fókuszában éppen a fenti kockázatok feltárása, kezelése, azokat lefedő tartalékok/tőke meghatározása áll. Az Európai Unióban az. u.n „single rulebook” [1] foglalja össze ennek a legfontosabb elemeit. Az u.n. baseli három pillér (minimális szabályozói tőkekövetelmény – szükséges gazdasági tőke meghatározása – nyilvánosság) rendszerében a hitelkockázatokra képzendő tőke meghatározására három lehetősége van a banknak: a standard módszer, az alap- valamint a fejlett belső minősítésen alapuló rendszer (FIRB. AIRB). Ezen belső minősítésen alapuló rendszerek részei a hitelminősítési rendszerek.

Egy IRB rendszer esetén a tőkeszámítás alapja a várható veszteség (EL – Expected Loss), amit a nemteljesítési valószínűség (PD – Probability of Default), a nemteljesítéskori veszteségráta (LGD: Loss Given Default) és a nemteljesítéskor az ügyféllel szemben fennálló követelés (EAD: Exposure at Default) szorzatából számolnak ki:

$$EL = PD * LGD * EAD$$

[1] Ennek elemei a CRR rendelet és a CRD IV. direktíva (EU Directive 2013/36/EU (CRD IV), EU Regulation 575/2013 (CRR) és ezek módosításai, CRR II (2019/876 EU Reg.), CRD V (2019/878 Dir.)), az egységes szanálási mechanizmus, a betétbiztosítási rendszer, az Európai Bankhatóság által kiadott sztenderdek (Reugalory Technical Standards, Implementing Technical Standards), valamint iránymutatások (guidelines), ajánlások (recommendation).

Hogy mi számít nemteljesítőnek, vagyis „default”-nak, azt szintén meghatározzák a szabályok[2]. A CRR 178. szakasza szerint:

„Egy adott ügyfél nemteljesítését akkor kell megtörténtnek tekinteni, ha a következők közül valamelyik vagy mindkettő bekövetkezik:

- az intézmény úgy véli, hogy az ügyfél valószínűsíthetően nem fogja teljes egészében teljesíteni hitelkötelezettségeit (…)
- a ügyfeleknek az intézménnyel, anyavállalattal vagy bármely leányvállalatával szembeni jelentős hitelkötelezettsége 90 napon túl késedelmes.”

A hitelintézeteknek tehát az EU-n belül egységes szabályok szerint kell default-nak besorolni egy-egy ügyletet.

PD, vagyis a „probability of default” annak a valószínűsége, hogy a következő egy éven belül beáll az adott ügyfélnél a nemteljesítés.

Az LGD az ügyfél nemteljesítésből származó veszteségnek és hitelezőnek az ügyféllel szemben fennálló, a nemteljesítés beálltakori kitettségének az aránya. Ez a behajtás sikerességétől és az esetleges biztosítékok meglététől függ. Amennyiben egy hitelintézet az IRB szerint szeretné a tőkekövetelményt meghatározni, úgy ezeket a számokat kell megbízhatóan megbecsülnie. Ehhez elengedhetetlen a minősítő rendszerek használata.

A „minősítő rendszer egy olyan szakértői rendszer, amely megvizsgálja az adott kérelmezőt, valamilyen belső modellre alapozva összehasonlítja az adatbázisában található

[2] A CRR 178. szakasza és a 2018/171. rendelet mellett az EBA DoD (azaz „Definition of Default”) standardja emelendő ki.

mintával és dönt arról, hogy az adott ügyfél kapjon-e hitelt" (KOVÁCS – MARSI: 84. o.). Ezen minősítő rendszereknek alapvetően két típusát különböztetjük meg: a rating és a scoring rendszereket. A ***rating rendszert*** akkor használják, amikor statisztikai módszerekkel nem vagy korlátozottan lehet éni, mert a kockázatok, ügyletek, ügyfelek egyediek, kisebb számúak de nagyobb értékűek. Ekkor a döntési kritériumok komplexebbek és nem mellőzik a szubjektív elemeket sem (pl. az adós stratégiája, jövőbeni piaci helyzete értékelése, vagy a management színvonala megítélése, stb.). A rendszer alapvetően az adós megítésésére fókuszál, maga az ügylet paraméterei mellékesek. Jellemzően a közepes és nagyvállalati hitelek és testre szabott, kis számban értékesített hiteltermékek esetén használatos (corporate ügyletek/ügyfelek). Az ilyen minősítő rendszerek outputja sokszor egy limit, nem is feltétlenül egy ügylet befogadása vagy elutasítása.

Ezzel szemben a tömegesen előforduló, standardizált, relatíve kis értékű ügyletek, melyek elsősorban a lakosság vagy mikro-, kis- és középvállalatok vesznek igénybe (retail ügyletek), ahol a statisztikai módszerek jól használhatók. Ezek egyszerűbb, objektív döntési kritériumokat használnak, de sok esetben nemcsak az ügyfél, hanem az ügylet jellemzőit is figyelembe veszik. Ezek a ***scoring*** rendszerek, melyek outputja egy-egy konkrét ügylet jóváhagyása vagy elutasítása.

A scoring rendszereknek is alapvetőn két fajtáját különböztetik meg. A jelentkezési scoringnak híjuk azt az esetet, amikor a hitelező egy új ügyfél kérelméről kell döntsön és nem áll rendelkezésére hiteltörténet vagy egyéb értékelhető belső információ (pl. számlatörténet, jövedelmi- költési szokások). Ekkor sokkal limitáltabb, általában az ügyfél által rendelkezésre bocsátott szociodemografikus adatokra (jövedelem, foglalkoztatottság,

lakhatás, családi állapot, stb.) és legfeljebb a hitelezők számára elérhető forrásokra (pl. Magyarországon a Központi Hitelnyilvántartó Rendszer) támaszkodhatnak.

A szabályozói/felügyleti szabályok a minősítő rendszerekkel szembeni elvárásokat is rögzítik. Ezek kiterjednek mind a kezdeti (kérelmi) minősítésre, mind az ügyfelek rendszeres (alapvetően évenkénti) felülvizsgálatára vonatkozóan. A részletekbe nem bocsátkozva általánosságban azt mondhatjuk, hogy a rendszereknek elő kell segíteni:

„a) az ügyfél, partner fizetőképességének, hitelképességének a kockázatvállalás előtti megállapítását,

b) az ügyfél, partner fizetőképességében, hitelképességében bekövetkező változások becslését, jövőbeni fizetőképességének megállapítását,

c) az ügyfél, partner nem teljesítési valószínűségének megállapítását, a nem teljesítési valószínűség számszerűsítését,

d) az ügyfelek, partnerek azonos kockázatokat és nem teljesítési valószínűségeket tükröző kategóriákba sorolását, a nem teljesítő ügyfelek, partnerek elkülönítését”[3],

valamint, hogy erre modellek is alkalmazhatók.

A hitelintézet maga döntheti el, hogy konkrétan milyen statisztikai/matematikai modell vagy szakértői eljárást használ a modell kialakításakor, „csak” a keretszabályoknak és a validációs szabályoknak kell megfelelni. Erre nézve részletes útmutatók, segédletek születtek[4]. A validációt a modell fejlesztésétől független személynek kell megtennie, annak

[3] 40/2016. (X. 11.) MNB rendelet az ügyfél- és partnerminősítés, valamint a fedezetértékelés prudenciális követelményeiről, 4. §

[4] Lásd pl. már 2008-ban a PSZAF validációs kézikönyvét.

ki kell terjednie a modell inputjainak, a modell kialakításának, a modell outputjainak, valamint korlátainak és teljesítményének értékelésére. A validáció minden fontos lépését dokumentálni kell, végül magát a validációt és egy független személynek értékelni kell.[5] A validációs eljárást rendszeresen, legalább 3 évente meg kell ismételni.

Amennyiben egy hitelintézet valamilyen modell alapján kívánja a tőkekövetelményt számolni, úgy ezt a belső validációs eljárás után, a SREP keretében a felügyelet ezt is ellenőrzi.

2.2 SZOKÁSOS FEJLESZTÉSI FOLYAMAT

2.2.1 Adatgyűjtés, mintavétel, definíciók

A fejlesztési folyamat előzményeként meg kell határoznunk, hogy mit tekintünk nemteljesítésnek. Ez a fentiek szellemében (DoD és egyéb szabályok) nem tűnik első látásra nehéz feladatnak, igazodhatunk az előírásokhoz és a belső szabályokhoz. De ez csak a PD becslésnél szigorú szabály, a bank a saját kockázati politikájából kiindulva választhat más kritériumot is. Amennyiben az így meghatározott deafult arány alacsony (kiegyensúlyozatlanság), ennek kezelésére is szóba jöhet ennél szigorúbb szabály is.

Következő kérdés az eredményidőszak meghatározása (lásd KOVÁCS – MARSI 121. o.): ez az az időszak, amíg a nemteljesítés bekövetkezését figyelembe veszi a modell. Ez a PD

[5] Lásd pl. a Magyar Nemzeti Bank 11/2022. (VIII.2.) számú „a hitelkockázat vállalásáról, méréséről, kezeléséről és kontrolljáról” szóló ajánlásának 174-176. pontjait.

becslésnél alapértelmezésben 12 hónap[6], de ettől a scoring modell kialakításánál el lehet térni.

Mivel a múlt adatait dolgozzuk fel, és ebből következtetünk a jövőre, szintén releváns kérdés, hogy milyen időtávot dolgozunk fel, vagyis hogyan határozzuk meg a mintavételi ablakot. Jelentkezési scoringnál nincs az ügyfélnek előélete, tehát nincs a „viselkedési ablak", de a "teljesítmény ablak", vagyis azon időszak meghatározása, amin belül figyeljük a nemteljesítést, releváns. Figyelni kell arra, hogy amennyiben minimum 90 napos késedelmet tekintünk default kritériumnak, ki kell zárni az ügyfeleket a mintából, ahol még ez nem telt el. Emellett javasolt a futamidő első néhány hónapjában nemteljesítővé vált ügyfelek kizárása is, mert ezek nagy része a tapasztalatok szerint csalás, ami nem hitelezési kockázati, hanem működési kockázati kategória, és szűrése a scoringgal nem hatékony, arra egyéb módszereket kell használni (ellenőrzések, feketelisták, stb.).

Törekedni kell a lehetőleg minél nagyobb elemszámra; néhány ezernél kisebb megfigyelésből valószínűleg nem fogunk tudni hatékony modellt összeállítani. Ha nem a teljes rendelkezésre álló adatmennyiséget dolgozzuk fel, vagyis mintát veszünk, akkor mindenképpen véletlenszerű legyen, esetleg rétegzett mintavétellel közelíthetjük meg a valós eloszlásokat (FAJSZI – CSER - FEHÉR).

Részben már említettük fent, hogy a kiinduló adatok összeállítása során figyelembe kell venni a teljesítmény ablakot, valamint célszerű kizárni a csalásgyanús ügyeleteket. Ezen kívül igyekezni kell a releváns információk összegyűjtésre, vagyis olyan adatbázis összeállítására, ami a lehető legjobban hasonlít a jövőbeni, várt ügyfelekre. Lehetőség szerint ki kell zárni ezért pl. a már nem forgalmazott termékeket, vagy olyan földrajzi

[6] Az értékvesztés képzésénél az IFRS 9 standard viszont a stage 2-ben lévő ügyleteknél már „lifetime" becslést ír elő.

régiókat, ahol már nem tervezünk értékesíteni (pl. fiókbezárás okán). Bizonyos esetekben (pl. hitelkártyák) az inaktív ügyfeleket is célszerű elhagyni, mert torzításhoz vezethet.

Általánosságban törekedni kell, hogy megbízható, ellenőrzött, hibátlan adatok kerüljenek be az adatbázisba.

A „kiegyensúlyozatlanság” problematikája, vagyis hogy a banki adatbázisok rendszerint csak azokat tartalmazzák, akik kaptak hitelt, és a rossz adósok aránya általában igen alacsony, szintén ebbe a témakörbe tartozik. Ezzel a 8. fejezetben foglalkozom még.

2.2.2 Adatelőkészítés

Az adtok előkészítésének egyik kiemelt terülte az adathiányok kezelése. Ezzel külön, az 5. fejezeben foglalkozom.

Az adatok előkészítéséhez tartozik a kiugró értékek, az outiler-ek kezelése. Egyértelmű definíció nincs arra, mit tekintünk outlier-nek, ez az konkrét adattól is függ. Nyilván egészen más egy kiugró érték pl. a gyerekszám, életkor vagy a jövedelem esetében. Célszerű a szórást, illetve annak egy sokszorosát (pl kétszeresét), és adott érték gyakoriságát figyelembe venni annak megítélésekor, valóban kiugró értékről van-e szó.

Ettől meg kell különböztetni azt az esetet, amikor ugyan van adat, de az nem érvényes (pl. negatív életkor vagy jövedelem a fenti példáknál maradva). Ezt hiányzó adatként kell értelmezni és kezelni.

2.2.3 Feltáró adatelemzés

Ezután következik az egyes változók egyenkénti vizsgálata. Először is meg kell értsük ezek jelentését, definícióját. Fontos megismerni karakterisztikájukat, terjedelmüket,

értékkészletüket, szórásukat, stb., és általában a pontosságukat, megbízhatóságukat. Számos más fent említett probléma (adathiányok, hibás adatrögzítés, stb.) ekkor kerül csak felszínre.

Leginkább a magyarázó erőt kell feltérképezni, vagyis hogy az adott változó menyiben „felel” a célváltozó alakulásáért. Erre számos módszer és mutató létezik, melyekről még a 6. fejezetben szólni fogok.

Ekkor szokott sor kerülni – amennyiben ez szükséges – a binnelésre vagy egyéb adatátalakításra is, amennyiben az szükséges; pl a kategórikus változók dummy változóra alakítása, ha regressziót kívánunk használni, vagy a standardizálásra, ha a az egyes változók nagyságrendje nagyon eltérő.

Fontos feladat, hogy a változók közti összefüggéseket feltárjuk. Ahogy majd az egyes modellek tárgyalásakor látjuk, erre több is érzékeny; az egymással erősen korreláló változók által reprezentált hatás is többszörösen is érvényesül, és torzíthatja modellünket.

Az összefüggések és a magyarázó erő feltárása után történhet a változószelekció, vagyis annak eldöntése, hogy mely változók kerüljenek be a modellünkbe. Ennek indokairól, módszereiről a 6. fejezetben szólok részletesebben.

2.2.4 A modell összeállítása

Maga a modell összeállítása rendszerint jóval kevesebb időbe és erőfeszítésbe kerül, mint az megelőző szakaszok. Kijelöljük a célváltozót, szétválasztjuk a tanító és tesztelő (validáló) adatbázist, majd a tanító adatbázison betanítjuk a prediktív modellünket. Majd azt a tesztkörnyezetben is futtatjuk, és összevetjük a valós értékekkel.

Rendszerint nem egyetlen modellt építünk, hanem többfélét is, vagy ha egy metodikában is gondolkodunk, rendszerint a változók kiválasztásához, valamely átalakítási lépéshez visszalépve próbálhatjuk annak teljesítményét javítani.

2.2.5 Modell kiértékelése

A modell kiértékelését mindig a teszt adatbázison kell elvégezni, mert a tanuló adatokon a modell általában lényegesen jobb teljesítményt mutat. Lehetséges módszereiről, mutatóiról betekintést az 7. fejezet ad majd.

2.2.6 Implementáció

Az implementáció kérdésével jelen dolgozatban nem foglalkozunk.

3 A FELADAT MEGHATÁROZÁSA

Konkretizáljuk, hogy ezen dolgozatban mi is a megoldandó feladat: adott egy finanszírozó, ahova a magánügyfelek nagy számban fordulnak kölcsönért. Ezen ügyfelek nagy része új ügyfél, nem rendelkezik hiteltörténettel az intézménynél, nincs tapasztalat viselkedését tekintve, nem ismerjük bankszámlaforgalmát, fizetési, költési szokásait és bevételeit. Azokra az információkra hagyatkozhatunk, melyeket az ügyféltől vagy más publikus információforrásokból begyűjthetünk. Feladatunk az, hogy megtámogassuk a pénzügyi intézmény döntését arról, hogy ügyletet jóváhagyja (folyósítsa a kölcsönt) vagy elutasítsa azt. A fentiek szerint tehát egy retail ügyfelekre vonatkozó kérelmi scoringot szeretnénk elkészíteni.

Itt egy predikciót kell tehát adnunk, múltbeli adatok elemzésével a jövőbeni kimenetet kell előre jeleznünk. Ez a jövőbeni viselkedés kétfajta lehet: az ügyfél visszafizeti a kölcsönét vagy sem. Kicsit egzaktabban fogalmazva: egy prediktív modellezést kell végrehajtani, azon belül is egy bináris osztályozást.

Olyan adatokat tudunk az elemzés körébe bevonni, melyek a potenciális ügyfelektől/ügyfelekről is begyűjthetők és az elbíráláskor már rendelkezésre állnak. További fontos követelmény, hogy a lényeges, a döntést valóban befolyásoló információkat gyűjtsük be. Egyrészt azért, hogy a modell áttekinthető maradjon, másrészt az információk begyűjtésének könnyítése érdekében. (Abban a tekintetben is komoly verseny van az egyes hitelezők között, hogy milyen egyszerű a hiteligénylés folyamata. Ráadásul a több begyűjtendő adat gyakran azzal a nem kívánt hatással jár, hogy az adatminőség romlik, több az adathiány vagy a hibásan megadott érték.)

4 LEGGYAKORIBB OSZTÁLYOZÁSI TECHNIKÁK

A professzionális hitelezők a minősítési rendszereikhez számos adatelemzési technikát használnak. A leggyakoribbak a következők:

- lineáris valószínűségi modell (lineáris regresszió),
- logitmodellek (logisztikus regresszió),
- diszkriminanciaanalízis,
- döntési fák,
- k-adik „legközelebbi szomszéd”-módszer, fuzzy clustering
- Naive Bayes
- Support Vector Machines
- neurális hálók
- genetikus algoritmusok
- gradient boosting, adaptive boosting
- hibrid módszerek

Ezeket a technikákat több szempont alapján is csoportosítani lehet. Ilyen csoportosítás lehet hogy felügyelt (pl. döntési fák) vagy nem felügyelt tanulásról van e szó (pl. neurális hálók), vagy hogy rendelkezésre áll-e már a konkrét üzleti döntést, elemzést segítő tapasztalat (meglévő termék), vagyis paraméteres vagy nem paraméteres az elemzés.[7]

A legkorábbi és legegyszerűbb modellek a lineáris és logisztikus regresszió valamint a a diszkriminancia elemzés voltak (BODON: 126.o.), előbbi a mai napig népszerű és referenciapontként használatos más módszerekkel való összevetésben. Először ezt fogjuk

[7] Lásd ilyen felosztást: KISS (2003) 50. o.

áttekinteni, majd a döntési fákat, a diszkriminancia analízist, a k-Nearest Neighbor, a neurális hálók és az SVM módszert vesszük jobban szemügyre.

4.1 LINEÁRIS REGRESSZIÓ

A regressziószámítás során két vagy több változó között egy sztochasztikus kapcsolatot tételezünk fel, és ezt a kapcsolatot törekszünk leírni (HUNYADI – VITA: 571). Célunk az, hogy felírjuk azt a függvényt, ami megadja, hogy a magyarázóváltozók egyes értékeinél az eredményváltozó milyen értéket vesz fel. Ezt általában így írjuk fel:

$Y= \beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \cdots \beta_k * X_k + \varepsilon$

A Y a függő vagy eredményváltozó, ßi: i magyarázó változó súlya, Xi: i. magyarázó változó, e: hibatag.

A lineáris regresszióval kapcsolatos fontos előfeltételek (VARGHA: 16. o.):

- a függő Y változónk kvantitatív, ami lehetőség szerint normális eloszlású,
- az eredményváltozó csak lineáris kapcsolatban lehet a magyarázó változókkal,
- az X magyarázó változónk függetlenek és lehetőleg szintén normális eloszlásúak,
- lehetőleg az Xi-k és Y együtt többdimenziós normális eloszlásúak legyenek. Ez azt jelenti, hogy bármelyik változó eloszlása normális a többi változó minden értékmintázata mellett (ez biztosítja a lineáris kapcsolatot).
- az ε maradékváltozó vagy hibatag várható értéke 0, különböző X értékeihez tartozó eloszlásai azonos szórást mutassanak és a legyenek korrelálatlanok (HUNYADI – VITA: 574.o.).

A regressziós modell minőségét azzal mérjük, mennyire illeszkedik a modell ahhoz az adatbázishoz, melyen épült, ezt pedig a legkisebb négyzetes eltérés alapján számolják (Y becsült és valós érté különbségének a négyzete).

Amennyiben ezt a metodikát hitelbírálatra kívánjuk használni, úgy az eredményváltozónk két értéket vehet fel: 1 amennyiben az adós rossznak minősül és 0, amennyiben jónak. Ha a feltételek teljesülnek, úgy Y értéke az adott X értékeknél felfogható úgy is, mint a nemfizetés relatív valószínűsége. Ha megadjuk, hogy Y, vagyis a nemfizetés mely valószínűsége esetén utasítjuk már el a kérelmet (ezt szokás „cut off"-nak vagy vágási pontnak nevezni), akkor meghatározhatjuk az elutasítandó kérelmeket.

A modell feltételrendszere a gyakorlatban nagyon sokszor nem teljesíthető: sokszor sérül a többdimenziós normalitás követelménye; ekkor a linearitást és a normalitást az egyes változókra külön kell megvizsgálni (VARGHA: 31.o). Még gyakoribb, hogy a hibatag nem normális eloszlású és varianciája nem állandó, tehát nem érvényesül a szóráshomogenitás. Ennek következményeként egyrészt a becsléshatékonysága csökken, és a regressziós együtthatók becsült varianciái és kovarianciái torzítottak és inkonzisztensek, így a szokásos tesztek (t- és F-próbák) nem érvényesek (KISS: 51. o.).

További probléma, hogy bizonyos X értékek esetében a becsült nemfizetési valószínűség értéke kívül eshet az értelmes 0-1 intervallumon. Igy a lineáris regressziót egyre kevésbé alkalmazzák (ORAVECZ 2007: 38. o.).

4.2 LOGISZTIKUS REGRESSZIÓ

A fenti hibák illetve korlátok egy részét a logit transzformációval ki lehet küszöbölni.

Ekkor függő változó helyett a függő változó egyik kategóriája – pozitív esemény (esetünkben a nemfizetés) – bekövetkezési valószínűségének (p), valamint a másik kategória – negatív esemény – bekövetkezési valószínűségének (1-p) a hányadosát, az u.n. odds-ot.

$$\text{odds}=\frac{P\ rossz}{P\ jó}=\frac{Prossz}{1-Prossz}$$

Az odds megmutatja, hogy mennyivel nagyobb az esélye, hogy a kérelmező negatív elbírálást kap, mint annak, hogy nem. Az odds már a [0,∞) intervallumon értelmezhető. Amennyiben ennek logaritmusát vesszük, akkor kapjuk meg az u.n. logit mutatót. A logit mutató már a (-∞,∞) intervallumon értelmezhető, illetve szimmetrikusabbá is tettük így a következők teljesülnek:

- a becsült érték a [0,1] tartományban marad
- a széleken nem nő/csökken túl gyorsan a becsült valószínűség értéke, mint ahogy az a lineáris regresszióval történő becslésnél előfordul.

Az illesztést a maximum likelihood módszerével végzik ebben az esetben (JÁNOSA: 284.o.).

Természetesen itt is szükséges a vágási értéket (cut-off) meghatározni, ami itt a rossz osztályba való besorolás esélyét mutatják.

A logisztikus regressziónak is további előnyei a lineáris regresszióval szemben:

- nem feltétel a változók normalitása
- folytonos és diszkrét változók egyaránt lehetségesek
- nem érzékeny a nemfizetők / fizetők arányára
- a ß-k a változók fontosságát is jelzik, nemcsak a súlyukat
- az eredmény a nem fizetés valószínűségét is megadja, ami mint fent láttuk szabályozói követelmény

Több hátrány azonban még mindig jelentkezik:

- nominális (minőségi) változókat nem képes kezelni
- az eljárás érzékeny a kiugró, extrém értékekre, az „outlier”-ekre
- nem robosztus adathiányossággal szemben sem

- a magyarázó változók közötti korreláció továbbra is rontja a modell hatékonyságát
- a nem lineáris összefüggéseket nem képes automatikusan kezelni ez a modell, ezért a magyarázóváltozó és a célváltozó közötti nem lineáris összefüggés esetén a magyarázó változó átalakítása szükséges.

Az első probléma viszonylag könnyen megoldható az. u.n. dummy változók bevezetésével. Ez azt jelenti, hogy amennyiben a szóban forgó minőségi változónak m számú az értékkészlete, akkor ezt a változót m-1 számú 0/1 változóval helyettesítjük, ahol mindegyik új változó egy-egy kategóriát képvisel az eredeti kategóriából 0 (nem) vagy 1 (igen) értékkel (HUNYADI – VITA: 667.o.).

Az outlierek kezelését az adatok előkészítése során kezelni szükséges. Az adathiány jellegétől és arányától függően lehet a hiányzó adatokat pótolni vagy elvetni (lásd részletesesebben a … fejezetet). A kiugró értékek kiszűrése ajánlott.

A multikorrealitás problémáját a változók szelekciójával lehet csökkenteni. Erről bővebben szólunk még a 6. fejezetben.

4.3 DÖNTÉSI FÁK

A döntési fa a tanuló algoritmusok körébe tartozó osztályozási eljárás, vagyis az egyedek besorolása leíró változóik (tulajdonságaik) alapján előre ismert osztályokba. A döntési fa, amit az eljárás alapján „Recursive Partitioning Algorithm”-nak azaz RPA-nak is szokás nevezni, egy gráf, amely különféle döntési eshetőségeket ábrázol. Az algoritmus elindul a teljes sokaságból (gyökér, root), és meghatározott, hierarchiába rendezett szabályok szerint részekre bontja a megfigyelések összességét. A részekre bontás célja, hogy a megfigyelések összességéből olyan csoportokat hozzon létre, amelyek a célváltozó kategóriáira nézve minél inkább homogének.

A legfontosabb, vagyis az egyedeket leginkább megkülönböztető, a célosztályt legjobban meghatározó változókat a fa a gyökér közelében teszteli, azokkal hasít, kiválasztva annak a legoptimálisabb értékét. Ezután az adatokat a fenti attribútum különböző értékei szerint osztályokba sorolja, ez a fa első elágazása. Az algoritmus ezután folytatja a módszert minden csúcsban, kiválasztva a legjobb osztályozó tulajdonságot és annak vágási értékét, amíg egy leállási kritériumot el nem ér (ekkor jutunk el az ágakon keresztül a levélhez, vagyis a nood-hoz). Ez akkor következhet be, ha:

- egy csúcson minden vagy majdnem minden adat egy osztályba sorolódik
- nem maradt olyan attribútum, amely alapján folytatni lehetne a hasítást
- valamilyen – általunk megadott - leállási szabály be nem áll.

A szétválasztásra, illetve annak jósága mérésére különböző mérőszámokat használnak az egyes algoritmusok (pl. $\chi 2$-statisztika – ez utóbbit használja például az egyik legelterjedtebb, a CHAID algoritmus – vagy valamilyen entrópiaindex (az ID3 algoritmus-család is ilyet használ), a Kolmogorov–Szmirnov-statisztika, C&RT, ami egy regressziós hasítási mérték, a szennyezettségi index, Gini-mutató (CART-család), stb.)[8].

Az, hogy az egyes levelek a célváltozó szempontjából melyik kategóriába esnek – ami a mi esetünkben azt jelenti, hogy nem fizetők lesznek-e vagy fizetők, úgy dől el, hogy az adott node-ban elhelyezkedő egyedek többsége melyik tulajdonsággal rendelkezik. [9]

Egy döntési fa addig nőhet, ameddig minden osztási kritérium elfogy és minden adat pontosan a helyére kerül. Azonban, ha a fa túl nagy, akkor számos döntés túl specifikus lesz és a modell túldefiniált lesz (overfitted). Ezért célszerű leállási szabályokat definiálni

---

8 Részeletesebben lásd: BODON: 143. o., JÁNOSA: 260. o.

9 Hasonló módon belső csomópontokhoz is rendelhetünk döntést. (BODON: 142. o.)

(fa előzetes metszése), ami megelőzi a túltanulást, ezzel csökkentse a méretet és javítja a szabályrendszer a hatékonyságát a későbbi adatok klasszifikációjára.

Az előmetszés (pre-prunnig) történhet úgy, hogy ha a fa növekedését egy előre magadott döntésszámnál, vagy osztályok minimális elemszámának előírásával szabályozzuk.

Alternatíva a post-pruning, ami a megnőtt fa visszametszését jelenti. A metszési kritérium a döntési pontokban számított hibaarányon alapszik. Ez hatékonyabb, mint az elő-metszés, mert lehetővé teszi, hogy az algoritmus meggyőződjön arról, hogy minden fontos adatmintát megtalált.

Amennyiben a modell a tanuló adatbázison kialakította a szabályokat, ezen szabályrendszer alapján tudja besorolni, becsülni később a teszt vagy tényleges adatokon a besorolást.

A döntési fa modell számos jó tulajdonsággal rendelkezik:

- Talán a legnagyobb előny, hogy a döntési szabályrendszer automatikusan kezeli – figyelmen kívül hagyja – az irreleváns változókat, ezért az adatelőkészítés során nem szükséges ezzel foglalkozni. A magyarázó változók kiszűrése azonban a fa mérete miatt hasznos lehet.
- Nem érzékeny a hiányzó értékekre, ezért a hiányzó értékek pótlására nincsen szükség az adatelőkészítés során és az összes rekord – az is, amely hiányzó értéket tartalmaz – használható a fa elkészítéséhez.
- Kategorikus (minőségi) változókat is kezel, nincs szükség transzformációra, dummy-vá alakításra
- Nem érzékeny az outlier-ekre
- Nem feltétel a normál eloszlás

- Nemcsak lineáris összefüggés lehet a magyarázó és az eredményváltozó között
- Az eredmény könnyen érthető, transzparens és jól indokolható laikusok számára is, megkönnyítve az üzleti bevezetést (vagy esetünkben egy ügyfél elutasításának okait is jól megmutatja), feltárja az összefüggéseket és az egyes változók magyarázó erejéről is képet kapunk
- Az egyes node-okban az eredményváltozó egyes kimenetei arányát is láthatjuk; ez azonban nem olyan jól és direktben használható, mint a regressziós modell PD-je

Néhány hátrányos tulajdonság:

- Problémát okozhat a magyarázó változók közti erős korreláció, vagyis vagy el kell hagyni, vagy esetleg össze kell vonni néhányat
- Könnyen kezelhetetlen méretűvé válhat
- Hajlamos a túltanulásra
- Noha az osztályba sorolás általában megbízható, nincs lehetőség a cutoff érték finomhangolására

Talán itt érdemes megemlíteni a hitelbírálati feladatokhoz szintén használatos **Random Forest** eljárást is, ami több döntési fa eredményét összesíti, növelve ezáltal a modell robosztusságát. Ezt úgy éri el, hogy nemcsak egy döntési fát állít fel, hanem egyszerre többet, oly módon, hogy a rendelkezésre álló adatállománynak és a változóknak is csupán egy véletlenszerűen kiválasztott részhalmazát használja fel, majd az így kapott modellek előrejelzéseit aggregálja (NYITRAI: 186. o.).

A túltanulási veszély csökkenése és a pontosság növekedése mint előny mellett hátrány a nehezebb átláthatóság és az algoritmus lassúsága, számolási kapacitásigénye.

4.4 DISZKRIMINANCIA ANALÍZIS

A diszkriminancia analízis egy osztályozási eljárás, ami azt vizsgálja, hogy mely változók alapján különülnek el a csoportok leginkább (JÁNOSA: 253. o.). Ezt az eljárás úgy határozza meg, hogy a lehető legnagyobb különbséget hozza létre a két csoport között. (A külső eltérések maximumát – és azzal együtt a belső eltérések minimumát – keresi.)

Előállítja azt a diszkrimináló függvényt, ami az egyes ismérvek (magyarázó változók) olyan lineáris kombinációját képezi, ahol ez teljesül. A függvény segítségével az új megfigyelések is besorolhatók a csoportba – esetünkben a nemfizetők vagy a fizetők csoportjába. Ehhez még egy cutoff értéket kell meghatároznunk, ami a téves besorolásból eredő legkisebb költség alapján lehetséges (ORAVECZ (2017): 612. o.).

Az eljárás feltétele a magyarázó változók normális eloszlása és a csoportbeli kovariancia-mátrixok megegyezőségét, illetve a különféle jellemzők függetlenségét.

A metodika előnyei:

- Egyszerű kezelhetőség, értelmezhetőség.
- A végeredményként kapott kategorizálásnál az összes tulajdonságot figyelembe veszi.

Hátrányok:

- A magyarázó változók normális eloszlása és a variancia- és kovarancia mátrixok azonossága a csoportoknál erős követelmény, nagyon sokszor nem biztosítható
- Meg kell határoznunk a téves kategóriákba sorolás költségét.

4.5 KNN

A kNN azaz k Nearest Neighbor metódus azon az elgondoláson alapszik, hogy a hasonló attribútumú objektumok hasonló tulajdonságokkal rendelkeznek. Tehát ha egy új egyed célváltozójának értékét (esetünkben hogy nemfizető vagy fizető adós lesz-e) akarjuk megbecsülni, ahol n számú tulajdonságot (magyarázó változót) vizsgálunk, akkor azt vizsgálja az algoritmus, hogy ebben az n dimenziójú térben melyik (k számú) tanuló-ponthoz van a legközelebb. Azt, hogy az új, vizsgált egyed melyik kategóriába kerül a célváltozó szempontjából, az dönti el, hogy a k számú legközelebbi szomszéd többsége melyik kategóriába esek. (Célszerű tehát páratlan k-t választani.) A hasonlóság nagyságát az euklidészi távolságfüggvénnyel mérjük. Ez lényegében már egy klaszterezési eljárás, felügyelt tanulás.

A k legközelebbi szomszéd osztályozó egy érdekessége, hogy nincs modellje. Minden egyes predikciónál ki kell számolni a páronkénti távolságot a kérdéses egyed és az összes tanító adatbázisbéli elem között, hogy a legközelebbi szomszédokat megtaláljuk.

A kNN előnyei:

- Jó távolság metrika nagyon jó eredményt tud elérni
- Nem lineárisan összefüggéseket is jól kezel

A kNN hátrányai (BODON: 175.o. alapján):

- Kategorikus változókra nem értelmezhető a távolság, ezért azokat transzformálni kell
- Nagyon eltérő nagyságrendekkel rendelkező változók esetében a nagyok hajlamosak elnyomni a kicsiket, ezért azok standardizálása szükséges lehet

- Nincs modell, amit megmutatná a besorolási döntés „indokait”, az egyes változók közti összefüggéseket
- A modell a hajlamos a lokális optimumnál megállni, más-más kiinduló ponttal ezért eltérő eredményeket kapunk. Célszerű többször futtatni és ezek közül választani.
- Ha a tanító adatbázis nagyon nagy, lassú lehet a predikció.

4.6 NEURÁLIS HÁLÓK

Neurális hálózatnak nevezzük azt az információfeldolgozó eszközt, amely azonos, vagy hasonló típusú feldolgozást végző műveleti elemek, un. neuronok hálózatából áll, rendelkezik tanulási algoritmussal és a megtanult információ előhívási, ill. bemutatási képességével. A neurális hálókat klasszifikációs, vagy regressziós célú felügyelt gépi tanulásra használják.

Neurális hálókat az idegsejtek működését szimulálják. A perceptron a bemeneti csatornáin több input adat (változók) kerülnek be, amelyből a tanulási folyamat elején előre megadott súlyok segítségével kiszámítja a kimenő értéket, és ha az meghaladja az ingerküszöbértéket, akkor továbbküldi az információt, ami az egyetlen kimeneti csatornán megjelenik. A feldolgozó függvény (súlyok) a tanulási folyamat közben változnak, annak érdekében, hogy a kimeneten a kívánt érték minél jobb becslése álljon elő. Ez a függvény – a feladattól függően - lehet lineáris vagy valamilyen szigmoid fügvény (KISS: 63 o.).

Amennyiben több, egymással párhuzamosan működő perceptront illesztünk egy gráfba, és ezek közvetlenül állítják elő a kimenetet, akkor ezt egyrétegű neurális hálónak nevezzük. A gyakorlatban azonban legtöbbször többrétegű hálót alkalmaznak; a bementi réteg és a kimenti réteg elkülönül, és a kettő között un. rejtett rétegek helyezkedhetnek el. Az egyes rétegek csak a szomszédos rétegben lévő neuronoknak adnak át bemenő adatokat.

A neurális hálót a tanuló adathalmazon tanítjuk be, a betanítás célja a veszteség (loss, cost) – vagyis a becslés és a valóság közti eltérés - minimalizálása a súlyok megfelelő kiválasztásával. A veszteséget a feladattól függően többféle módon is meghatározhatjuk. Az egyik leggyakrabban használt veszteségfüggvény az átlagos négyzetes hiba (MSE, mean square error) amely az aktuálisan számolt és a tényleges értékek négyzetes különbségének átlaga.

A mesterséges neurális hálózatok fő sajátosságai:

- Legnagyobb előnye a tanulási képesség
- Nem lineáris összefüggéseket is kiválóan kezeli
- A besorolás eredményeképp az egyes kategóriákba kerülés valószínűségét is megismerhetjük
- Hajlamos a túlillesztésre. Ezt a tanulási folyamat megfelelő időben történő leállításával lehet kezelni.
- Hiányzó értékekre érzékeny. Kevés hiányzó érték esetében megoldás lehet a hiányzó értéket tartalmazó rekordok elhagyása. Amennyiben ez nem megoldható – nagyszámú hiányzó érték esetén –, akkor a hiányzó értékeket pótolni szükséges.
- Érzékeny a kiugró értékekre, ezért azokat ki kell szűrni az adatelőkészítés során.
- A legnagyobb hátránya az, hogy mintegy fekete dobozként működik. A felhasználó számára a besorolás kritériuma nem ismert, nem látjuk az összefüggéseket.

4.7 SVM

A korábbiakban többnyire olyan klasszifikációs eljárásokat vizsgáltunk, melyek a tér lineáris szeparációján alapultak, ez azonban csak nagyon ideális esetekben áll elő. A Support Vector Machines metodika ezt a problémát oldja meg a kernel módszerrel, aminek

a lényege, hogy egy adott (a változók száma által meghatározott) dimenziójú térben lineárisan nem szeparálható osztályok egy magasabb dimenzióban nagyobb valószínűséggel lesznek szeparálhatók.

A transzformáció után az algoritmus olyan hipersíkot keres, ami a legjobban választja el a két szétválasztandó csoportot. Ez elméletben az, ami a legtávolabb fekszik a két csoportból. gyakorlatban a legkisebb „költség" alapon határozzák ezt meg, vagyis miden hibához (a síkon belül került egyedhez) rendelnek egy költséget és az algoritmus célfüggvénye ennek a minimalizálása. Az SVM algoritmusok jellemzően a hiba nagyságának beállítását, valamint a kernel függvény formáját kérik a felhasználótól.

Az SVM tulajdonságai:

- Az SVM csak numerikus változókat tud kezelni
- A magyarázó változók eltérő nagyságrendi skálái zavaróak, ezért célszerű azokat standardizálni
- A teszt adatbázisnak ugyanolyan szerkezetűnek kell lennie, mint train adatbázisnak
- Kiváló klasszifikációs tulajdonsággal rendelkezik

4.8 HIBRID MÓDSZEREK

Az újabb fejlesztésű modellek (neurális hálók, SVM, genetikus módszerek) mellett az irodalomban és a különböző internetes forrásokban[10] egyre többször fordulnak elő az u.n. hibrid módszerek.

Az egyszerű hibrid modell esetében a modellezés különböző fázisaiban eltérő modelleket használnak, kihasználva azok erős tulajdonságait. Igy a változószelekció, a modell

[10] pl. összefoglalóan: LI: 185.o., MARKOV: 191.o.

paraméterei meghatározása és az osztályozás más modell segítségével végzik. Ilyen lehet pl. a változószelekció esetén a logisztikus regresszió használata, majd egy neurális háló vagy SVM az osztályozásra.

Némileg más az „ensemble learning” amikor több osztályozást végeznek különböző minták alapján, majd ezekből kiválasztják a legjobbakat és ezek közös eredményéből alakul ki a végső osztályozás.

5 ADATHIÁNYOK KEZELÉSE

Bármilyen adatelemzési feladatot hajtunk végre valós adatokon, szinte elkerülhetetlen, hogy ne találkozzunk a hiányzó adatokkal. Az adatelőkészítés során ezért a hiányzó adatok számát és mintázatát ellenőrizni kell.

Két fő hiányzó adattípust különböztetnek meg:

- MCAR: missing completely at random: az adathiány teljesen véletlenszerű, nem áll kapcsolatban sem a célváltozó sem az éppen vizsgált magyarázóváltozó vagy a többi változó értékétől.
- MNAR: missing not at random: A hiányzó adatok összefüggésben vannak a változó értékével.

Az MNAR esetében célszerű megvizsgálni az adatgyűjtés folyamatát és megkeresni az okokat. (pl. ha egy kérdőív kitöltésekor sokan nem válaszoltak egy adott kérdésre, vagy egy értékét szisztematikusan nem választják, inkább nem töltik azt ki – pl az eladósodott ügyfelek nem adják meg, hogy hol van máshol hitelük).

Nyilván az MCAR típusú adathiány a kisebb nehézség, de ha az adathiány nagy, akkor az is okozhat problémát. Nagy adathalmazoknál 5% az elfogadott határ. Ha ennél magasabb, akkor célszerű a változót kihagyni, ha az nem okozza a modell hatékonyságának jelentős csökkenését.

Ha a hiány mértéke nem túl magas, akkor a hiányos adatokkal rendelkező megfigyelések teljes törlése a legegyszerűbb megoldás. Ez azonban csak MCAR adathiány esetében járható út, különben torzításokat okoz.

Az is lehetséges, hogy az egyes változók elemzésekor az adott változó összes értékét vizsgáljuk, függetlenül attól, hogy valahol egy másik változónál adathiány van. Ez azonban több változós elemzésnél nehezen kivitelezhető.

A probléma kezelésének gyökeresen más módja a hiányzó adatok helyettesítése, imputálása. Erre több módszer is létezik. A legegyszerűbb változat, ha a hiányzó adatokat az átlaggal vagy mediánnal, modusszal helyettesítjük; ez a középértéket nem változtatja, de a szórást igen. Finomíthatjuk ezt úgy, hogy ezt kisebb csoportokra bontva végezzük. Vannak ennél sokkal kifinomultabb technikák az imputálásra. Lehet a többi változó alapján egy regresszió segítségével megbecsülni a hiányzó értéket. Megjegyezném, hogy a kategórikus változók esetében az imputációt nem javasolják (ha mégis szükséges, akkor a modusszal helyettesítés jöhet szóba).

A „Hot deck" imputáció esetén a leginkább hasonló hiánytalan esetet keresik meg és annak értékét veszik át, ez lehet véletlen választás vagy valamilyen távolságbecsléssel keresik meg a leginkább hasonló esetet és annak értékével pótolják a hiányzót, esetleg több ilyenből véletlenszerűen választanak.[11]

„Cold-deck" imputáció esetén külső forrásból (pl. múltbéli esetek) próbálják a hiányzó értéket megbecsülni.

A többszörös imputáció esetén a bizonytalanságot és véletlenszerűséget próbálják visszahozni az adatokba azáltal, hogy a többszörös értéket választanak. Ezzel az adatbázis is megtöbbszöröződik, az elemzés is nehezebbé válik, de bizonyos szoftverek már segítenek ezen nehézségek leküzdésében.

---

[11] Részeletesebben ír az imputációs eljárásokról pl: HUNYADI – VITA: 302 o.; ORAVECZ (2008): 22. o.

A fenti módszerek a MAR adathiányok kezelésére alkalmasak; léteznek már MNAR hiánykezelésre is módszerek (Lásd ORAVECZ (2008): 30-31. o.), de ezek leírástól most letekintünk.

6 VÁLTOZÓSZELEKCIÓS ELJÁRÁSOK

Elképzelhető, hogy egy-egy ügyfél vagy ügylet kockázatának megítéléshez nagyon sok információ áll rendelkezésünkre. De fordítva is megfogalmazható a kérdés: milyen kérdéseket tegyünk fel a potenciális ügyfelünknek, milyen információkat gyűjtsünk be? Melyek a kockázatot valóban, erőteljesen befolyásoló tulajdonságok? Ezek a kérdések elsősorban a modellek tervezésénél és későbbi fejlesztésénél vetődnek fel.

Azonban ahogy láttuk, a már rendelkezésre álló információk közül is több okból előnyös és szükséges is a változók szelekciója:

- A modellek többsége nem kezeli jól az egymással erősen korreláló változókat. Ha ilyenek maradnak a modell tanító adatbázisában, akkor az ronthatja a modell hatékonyságát, hiszen egyfajta hatás többszörösen is megjelenik.
- A „dimenziós átok” (curse of dimensionality) jelentkezése; ennek lényege, hogy ahogy a magyarázó változók száma, vagyis adatinkban a dimenziók száma emelkedik, azzal exponenciálisan nő a tér, amelyben a megfigyeléseket értelmezni szükséges, az adatok pedig egyre ritkábban töltik ki ezt a teret, egyre több üres pont lesz ebben a multidimenziós rácsban. Mindez nemcsak az elemzést teszi nehézzé, de a számítási kapacitást is növeli.
- Üzleti/folyamat- szempontból is előnyös, ha minél kevesebb változót használunk: egyrészt modellünk átláthatóbb lesz, ami elfogadottságát növelheti, másrészt kisebb erőfeszítésekkel vélhetően jobb adatminőségben, kevesebb hiányzó adattal rendelkező adatbázissal dolgozhatunk.

Mennyiségi változók között a kapcsolatot korrelációs vizsgálattal, vegyes kapcsolat esetén variancia-elemzéssel (ANOVA) tudjuk megvizsgálni.

Amennyiben azt tapasztaljuk, hogy egyes változók között az összefüggés magas, akkor érdemes az egyiket elhagyni, vagy egy mutatóval helyettesíteni azt. Pl. egy jelzálog-hiteleket vizsgáló elemzés során valószínűleg erős lesz a kapcsolat a hitelösszeg nagysága és a fedezetül szolgáló ingatlan piaci értéke között. Ha ezzel modellünk magyarázó ereje nem csorbul jelentősen, elképzelhető, hogy érdemes e kettő helyett ezek hányadosát, a fedezettségi mutatók használnunk.

A csekély magyarázó erővel bíró változók kiszűrésére legtöbbször a lift, a WoE, vagy a IV mutatót használják a gyakorlatban. A *lift* azt mutatja meg, hogy egy „kiválasztott *csoporton belül hányszor nagyobb a célesemény bekövetkezésének valószínűsége, mint a teljes mintában"* (FAJSZI – CSER – FEHÉR 116. o.). Pl. kiválasztunk egy három értékű kategórikus mutatót, és mindhárom ezáltal képzett csoportra meghatározzuk, hogy ott mekkora volt a nemfizetési arány. Ezt az arányt osztjuk az összes esetre számított nemfizetési aránnyal. Képlettel:

$$\text{LIFT} = \left(\frac{P\ rossz\ csoport}{P\ rossz\ teljes}\right)$$

Ha egy olyan változót találunk, ahol az egyes csoportok lift-értékei nem térnek el jelentősen egymástól (vagyis 1 környékén állnak), annak a változónak csekély a magyarázó értéke, elhagyható a modellből.

A *WOE* (Weight of evidence) esetében hasonlóan, a *célváltozó kategóriái relatív arányait, az odds-okat vizsgáljuk az egyes csoportban.* Esetünkben ez úgy néz ki, hogy a rossz adósok arányát osztjuk a jó adósok arányával és ennek veszük a logaritmusát.

$$\text{WOE} = \ln\left(\frac{P\ rossz}{P\ jó}\right)$$

A WOE használata nemcsak a változok magyarázó erejének áttekintéséhez hasznos, hanem pl. egy logisztikus regresszió esetén a kategórikus változókat is be tudjuk ennek segítségével vonni az elemzésbe. Ez az u.n. *WOE-transzformáció*, ami azt teszi, hogy a kategórikus változók egyes értékeire kiszámított WOE-értékre cseréli le az eredeti értéket.[12]

A IV (*Information Value*) is kedvelt módszer a magyarázó erő kimutatásához. Ezt a következőképp számolhatjuk ki (FAJSZI - CSER – FEHÉR: 362. o. alapján):

$$IV\ c = \Sigma (p_{rosszak\ i} / p_{jók\ i})^* WOE_i$$

ahol i a kiválasztott C változónk egyes osztályait jelöli, a WOE_i pedig az egyes osztályokra a fent bemutatott módon kiszámított WOE mutatót.

Természetesen ennél jóval bonyolultabb változószelekciós eljárások is alkalmazhatók. A statisztikai programok ezeket tartalmazzák, így ezek matematikai hátterére szükségtelen kitérni.

A regressziószámításnál a legegyszerűbb a valamelyik lépésenkénti (stepwise) módszert beilleszteni, amikor az egyes változókat vagy előremenően (forward) vagy hátrafelé (backward) szelektáljuk. A backward módszer esetén a lehető legtágabb modellből indulnak ki, és azt szűkítik lépésről lépesre, minden lépésben egy változót kiejtve addig, amíg a meghatározott egy szempont szerint elfogadható, de legegyszerűbb (legkevesebb változót tartalmazó) modellig sikerül eljutni.

---

12 Részletesebb leírást lásd pl. FAJSZI – CSER – FEHÉR: 127-128 . o. illetve 363-365. o.

A forward módszer ennek a fordítottja, a legnagyobb magyarázó erővel bíró változóval indulva addig bővíti a modellt, amíg a magyarázó erő szignifikánsan nő.

A „best subset" esetén pedig iterációval keresik a legjobb változó-variációt.

A magyarázó erőt illetve a modell „elfogadhatóságát" többféleképpen is lehet mérni: ez lehet a determinciós együttható, a Theil-féle szabadságfokkal korrigált determinációs együttható, az AIC (Akaike Information Criterion), esetleg más mutató (HUNYADI: 677-678. o.).

Léteznek emellett u.n. „zsugorító módszerek" is, melyek nem a változót zárják ki, hanem a béta esztimátorok méretét csökkentik úgy, hogy az információs veszteség ne legyen jelentős. Ilyen a „ridge" módszer vagy annak egy módosított változata a LASSO (least absolute shrinkage and selection operator).

7 KIÉRTÉKELÉSI ELJÁRÁSOK

A kialakított modellek teljesítményének mérése, kiértékelése és összehasonlítása fontos része a modell véglegesítésének és a rendszer kialakításának.

Mivel egy osztályozási feladatról van szó, ezért a legkézenfekvőbb a klasszifikációs mátrix elkészítése ez esetünkben – ahol a pozitív esete a nemfizetés - így néz ki:

	Predikció szerint pozitív	Predició szerint negatív
Pozitív (nemfizető)	TP	FN
Negatív (fizető)	FP	TN

1. sz. táblázat: saját összeállítás

Ha egy negatív esetet pozitívnak sorolunk be (FP), az az első fajú hiba. Ha egy negatívnak besorolt viszont valójában pozitív (FN), az a másodfajú hiba.

Ebből a besorolásból számos mutató képezhető, mely mind-mind más szempontból értékeli az eredményt:

- Accuracy: az összes helyesen besorolt aránya az összes esethez: (TP+TN)/ (TP+TN+FP+FN)
- Sensitivity (vagy recall): az összes pozitív esetből a helyesen felismert aránya: TP/(TP+FN)
- Precision: az összes pozitívként besorolt mekkora arányban volt valójában pozitív TP/(TP+FP)
- Specificity: a negatívok helyesen felismert aránya: TN/ (TN+FP)
- FNR: False negatív rate. FN/(TP+FN)
- FPR: False positive rate: FP/(FN+TP)

Könnyen belátható, hogy az egyes tévedéseknek nem egyforma a súlya, esetünkben a másodfajú hiba, a FN „fáj" a legjobban, hiszen ekkor egy olyan adóst hitelezünk meg, aki nemfizetővé válik és ebből veszteségünk származik. Természetesen tévesen elutasított kérelmekből (FP) is származik veszteség (helyesebben elvesztett haszon), hiszen ekkor a hitelező elesik a kamatkülönbözettől. A fenti „nyers" táblázatot ezért sokszor egy költségmátrix-szal szorozzák fel, és így számolják ki a tévedések valós költségét. Ettől a hagyományos felfogástól némileg eltérő költségmátrixot javasol Oravecz.[13]

A kiszámolt mutatókból is különös figyelemmel kell kövessük a Sensitivity-t és a FNR-t. Az első helyen szereplő és első látásra valóban a legfontosabbnak tűnő Accuracy azért lehet ennél a feladatnál félrevezető, mert a pozitív arány általában alacsony. Ezért ha egyszerűen minden ügyet negatívnak sorolunk be, akkor is viszonylag magas ennek a mutatónak az értéke.

Az osztályozási feladatok zöménél a ROC görbét is fel szokták rajzolni. Az X tengelyen a kumulált FP arány, míg az Y tengelyen a kumulált TP arány látható, vagyis azt tudjuk meg, hogy az ügyfelek valahány százalékának elutasításával milyen arányban utasítjuk el a nemfizetőket. Modellünk akkor teljesít jól, ha az meredeken emelkedik az Y tengely mentén. Az origóból kiinduló 45 fokos egyenes a véletlen besorolással egyenértékű.

Az összemérést segíti, ha a görbe alatti területet kiszámoljuk, ez az AUC vagy AUROC mutató. Alternatívaként a Gini- mutatót is használják, ami a ROC görbe és a véletlenszerű kiválasztást mutató egyenes kétszeresével egyenlő.

---

[13] A valós cash-flowból vezeti le az egyes döntésekhez rendelhető összeget, így a FN, vagyis a tévesen elutasított ügyletekből eredő elvesztett hasznot nem veszi figyelembe, viszont a helyesen befogadottakon keletkezőt igen (Oravecz 2007: 624-625. o.).

A szabályozók (lásd pl. PSZAF validációs kézikönyv 406. pontját) sem határoznak meg egy egyértelmű mérési módszert, a ROC görbe, az AUROC, Gini együttható, Accuracy Ratio, Mann-Whitney statisztika, Wilcoxon-Mann-Whitney statisztika, Kolmogorov-Smirnov (K-S) statisztika egyaránt említésre kerül.

8 PROBLÉMÁK A SCORING MODELL ÉPÍTÉSÉNÉL

Mint minden predikciós feladatnál, a legfontosabb probléma az, hogy a múltban megtörtént esetekből próbálunk a jövőre következtetni. Korábbi hitelek és hitelkérelmek adatot dolgozzuk fel, noha eközben a körülményekben változások állhatnak be. A legfontosabb, hogy az általános gazdasági helyzet, a konjuktúra-hullámok alapvetően befolyásolják az ügyfelek fizetőképességét. Ezt semmilyen modell nem képes előre jelezni; az ilyen negatív forgatókönyvekre elsősorban a tőketartalékoknak kel fedezetet biztosítani.

Szintén fontos külső körülmény a versenyhelyzet: a versenytársak befogadásában, kockázatvállalásában bekövetkező változások befolyásolják azt is, hogy hozzánk milyen ügyfelek/ügyletek érkeznek; ha ezek összetétele megváltozik, úgy a meglévő portfólióból eredő tapasztalatok sem konvertálhatók egy-az egyben az új helyzetre.

Sűrűn előforduló eset, hogy a bank új terméket vezet be, vagy egy új ügyfélkört céloz meg (pl. területileg, egy új fiók megnyitásával); ekkor is korlátozottan használhatók a múltból eredő adataik feldolgozásából kialakított modelljeink.

Ezen – a modelleken kívül eső körülményeken kívül – azonban vagy igen fontos, modellezési hiányosság/probléma is. Ez pedig az adataink rendszerinti kiegyensúlyozatlansága. Ha csak a meglévő ügyfeleink adatai dolgozzuk fel, akkor abban minden bizonnyal kis számban lesznek fellelhetők a rossz adósok. Ezért célszerű az elutasított kérelmek adatait is rögzíteni és bevonni modellbe (hiszen a jövőben is fogunk ilyenekkel találkozni); ez az u.n. reject inference (POLÁK – KOCSIS (2015): 22-23. o., ORAVECZ (2009): 63.o. alapján). Ezen esetekben becsülni kell az ügyfél magatartását, hogy az milyen lett volna, ha hitelt kap. Hitelezés esetében a „nyitott kapuk” módszere – amikor egy ideig mindenkit befogadnak, hogy ezen mintából utána következtetni tudjanak

– annak nagy költségei miatt nem járható út. A „résnyire nyitott kapu" több megoldást is magába foglal: lehet egy véletlenszerűen rövid időre bárkit befogadó minta gyűjtése, lehet egy nemfizetés esetén kisebb várható veszteséggel (LGD) rendelkező ügyfélkör befogadása. Ezek mellett léteznek statisztikai megoldások is az elutasított ügyfelek viselkedésének becslésére, mint az átsúlyozás, extrapoláció. [14]

[14] Részletesen lásd ORAVECZ (2009): 66-96 o.

9 EGY MINTA-MODELL KIALAKÍTÁSA R-BEN

A modell építése ismertetése előtt szeretnék néhány alapvető észrevételt tenni:

- Nem volt célom, hogy egy jól működő és megfelelő hatékonyságú modellt készítsek, ez messze meghaladná egy szakdolgozat kereteit. Márcsak a felhasznált adatbázis sem alkalmas ilyenre, hiszen az nem valós, minden bizonnyal torzított adatokat tartalmaz.
- A feljebb vázolt metodika – pl változószelekció, stb. - illetve a szóba jöhető statisztikai modellek közül is csak néhány illusztrálására törekedtem.
- Ezért a modell eredménye másodlagos volt, és az nem a módszerek hibája, sokkal inkább a fentiek és a szerző tapasztalatlanságának következménye.
- Csak az osztályozást végeztem el (nem fizető / fizető adós), nem készítettem PD becslést és LGD becslést.
- Két módszert – talán a leggyakoribbakat - használtam erre: logisztikus regressziót és döntési fát.

9.1 AZ ADATBÁZIS BEMUTATÁSA

Az elemzéshez használt adatbázist a Kaggle oldaláról töltöttem le (Kaggle „BLSD” https://www.kaggle.com/datasets/zaurbegiev/my-dataset). Olyan adatbázist kerestem, ami nagyszámú megfigyelést tartalmaz és a magánszemélyek adatait tartalmazza lehetőség szerint minél több szempontból. Cél volt az is, hogy viszonylag friss legyen, ne kapcsolódjon hozzá korábbi nagyszámú elemzés. Az adatok feltáró adatelemzését, az adatok előkészítését és a modell építését R-ben végeztem el.

Az adattábla 100.514 megfigyelést és 19 argumentumot tartalmaz. Szerkezete a következő:

```
 'data.frame':   100514 obs. of  19 variables:
$ Loan.ID                  : chr  "14dd8831-6af5-400b-83ec-68e61888a048" …
$ Customer.ID              : chr  "981165ec-3274-42f5-a3b4-d104041a9ca9"…
$ Loan.Status              : Factor w/ 2 levels "Charged Off",..: 2 2 2 2 2 1 2 1 2 2 ...
$ Current.Loan.Amount      : int  445412 262328 99999999 347666 176220 206602 217646…
$ Term                     : chr  "Short Term" "Short Term" "Short Term" "Long Term" ...
$ Credit.Score             : int  709 NA 741 721 NA 7290 730 NA 678 739 ...
$ Annual.Income            : int  1167493 NA 2231892 806949 NA 896857 1184194 NA 2559110…
$ Years.in.current.job     : chr  "8 years" "10+ years" "8 years" "3 years" ...
$ Home.Ownership           : chr  "Home Mortgage" "Home Mortgage"  "Own Home" ...
$ Purpose                  : chr  "Home Improvements" "Debt Consolidation" "Debt Consolidation" …
$ Monthly.Debt             : num  5215 33296 29201 8742 20640 ...
$ Years.of.Credit.History  : num  17.2 21.1 14.9 12 6.1 17.3 19.6 8.2 22.6 13.9 ...
$ Months.since.last.delinquent: int  NA 8 29 NA NA NA 10 8 33 NA ...
$ Number.of.Open.Accounts  : int  6 35 18 9 15 6 13 15 4 20 ...
$ Number.of.Credit.Problems : int  1 0 1 0 0 0 1 0 0 0 ...
$ Current.Credit.Balance   : int  228190 229976 297996 256329 253460 215308…
$ Maximum.Open.Credit      : int  416746 850784 750090 386958 427174 272448 272052…
$ Bankruptcies             : int  1 0 0 0 0 0 1 0 0 0 ...
$ Tax.Liens                : int  0 0 0 0 0 0 0 0 0 0 ...
```

2.sz. táblázat: saját forrás (R export)

Az első két oszlop a hitel és az ügyfél egyedi azonosítója. Ez számunkra nem hordoz információt, el elehet távolítani.

A harmadik oszlop, aminek két lehetséges értéke van (ezt az R unique parancsával kérhetjük le): „Fully Paid” és „Charged Off”, lesz a célváltozó. Azokat az egyedeket keressük, ahol a státusz ez utóbbi, vagyis leírt. (Ez lesz a pozitív esetünk.)

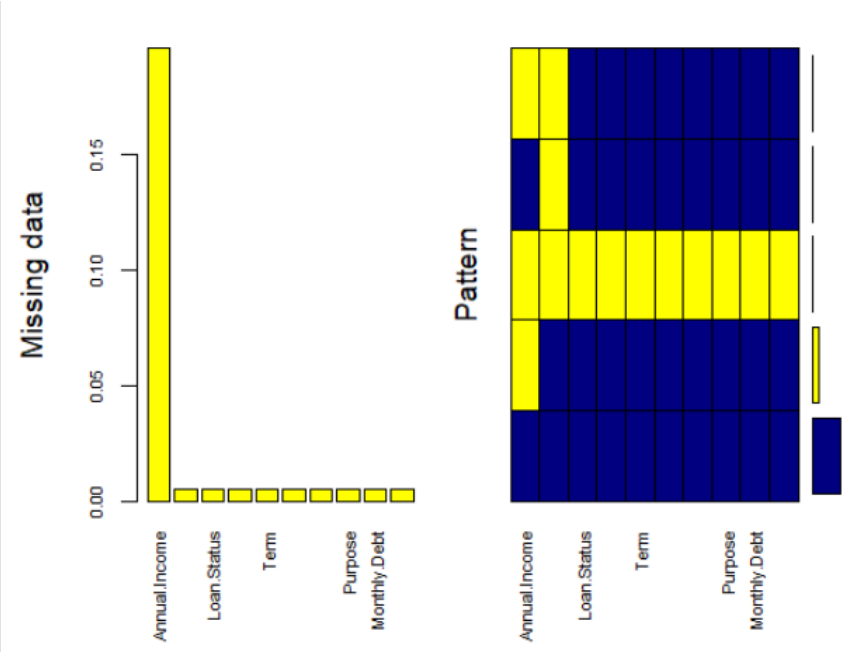
A hatodik változó a „Credit Score” egy külső minősítő minősítésének értékét tartalmazza. Miután mi pont ezt szeretnénk saját magunk megállapítani, ezt is figyelmen kívül fogjuk hagyni. A 13-15 oszlopok („Months.since.last.delinquent”, „Number.of.Open. Accounts” „Number.of.Credit.Problems”) olyan információkat tartalmaznak, melyek nem állnak rendelkezésre a hitelkérelem beadásakor, ezért ezeket is figyelmen kívül hagyjuk, hiszen célunk egy hitelkérelem beadásakor meghozandó döntés modellezése.

A fent említett oszlopok eltávolítása után 10 oszlopunk marad.

9.2 FELTÁRÓ ADATELEMZÉS, ADATOK ELŐKÉSZÍTÉSE

Mielőtt az egyes változókat egyesével megvizsgálnánk, a ***hiányzó adatokat*** vesszük górcső alá. Azonban nemcsak az adatbázisban „NA”-ként megjelölt sorokat kell figyelnünk, hanem azt is, hogy az adatok felvitelekor nem helyettesítették azt valamilyen más rövidítéssel. Az R-ban már az adtok beolvasásakor is meg lehet adni, hogy milyen karaktereket kezeljen hiányzó adatként.[15]

Ezután a „MICE” csomag segítségével felrajzoltam, hogy adathiányoknak milyen a térképe.



1. *sz. ábra: saját forrás (R export)*

Jól látszik, hogy van 514 olyan megfigyelésünk, ahol szinte minden adat hiányzik; ezekkel nem tudunk mit kezdeni, ki kell dobnunk az adatbázisból. Ezen kívül az Annual Income esetében van számottevő 18.343 hiányzó adatunk. Ezzel is kezdenünk kell valamit. A

[15] Én az első próba-beolvasás után az „ ”, „n/a”, „NA” karaktersorozatokat adtam meg és ezzel a rejtett hiányokat ki is lehetett szűrni.

gyakorlat szerint 5% felett esetén helyesebb a mutatót kihagyni a vizsgálatból, vagy valamilyen imputációs eljárást alkalmazni. Én most egy egyserű átlaggal helyettesítést használtam. Hiány még „Years in Current Job” esetében volt; mivel ez egy 10 db-os értékkészlettel rendelkező minőségi (kategorikus) változó, itt esetleg a modusszal való helyettesítés jöhetne szóba, de ezt a várható nagy torzítás miatt elvettem, a minta nagyságára tekintettel ezeket a rekordokat is töröltem az adatbázisból. Még mindig 95.776, már hiánytalan adatú egyeddel rendelkezünk, ami több mint elegendő.

Ezután megkezdhetjük az egyes változók megismerését. A következőkre fokuszálunk elsősorban: eloszlások, terjedelem, outlier-ek. Jelenleg ezek a változóink fő tulajdonságai:

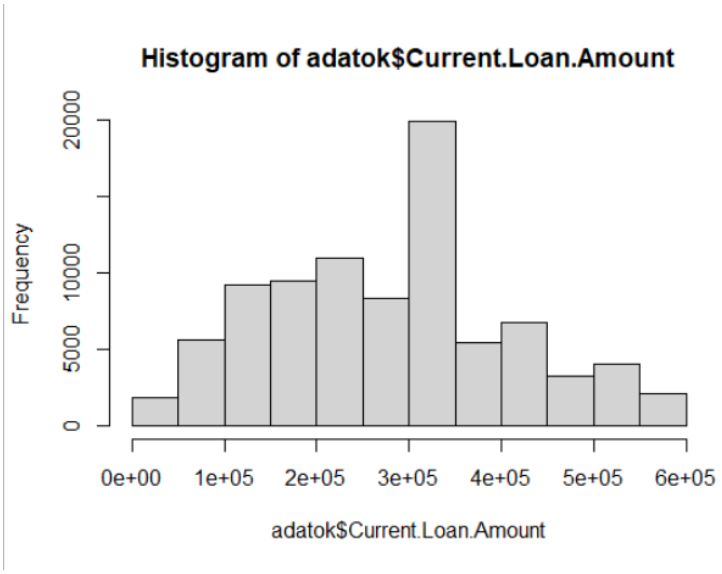
Loan.Status	Current.Loan.Amount	Term	Annual.Income	Years.in.current.job
Charged Off:21365	Min. : 10802	Length:95776	Min. : 76627	Length:95776
Fully Paid :74411	1st Qu.:183678	Class :character	1st Qu.: 942970	Class :character
	Median : 316954	Mode: character	Median : 1373246	Mode: character
	Mean :11822765		Mean : 1394126	
	3rd Qu.: 528836		3rd Qu.: 1530502	
	Max. :99999999		Max. :165557393	
Home.Ownership	**Purpose**	**Monthly.Debt**	**Years.of.Credit.History**	**Maximum.Open.Credit**
Length:95776	Length:95776	Min. : 0	Min. : 3.6	Min. :0.000e+00
Class :characte	Class :character	1st Qu.: 10442	1st Qu.:13.4	1st Qu.:2.762e+05
Mode: character	Mode :character	Median : 16435	Median :16.8	Median :4.713e+05
		Mean : 18703	Mean :18.0	Mean :7.640e+05
		3rd Qu.: 24255	3rd Qu.:21.5	3rd Qu.:7.870e+05
		Max. :435843	Max. :70.5	Max. :1.540e+09

2. sz táblázat: saját forrás (R export)

Elsőnek a *Current Credit Amount változót* vizsgáljuk meg: feltűnő, hogy a terjedelem nagy, a maximum nagyon magas 99.999.999. Egy kis kutakodás után kiderül, hogy az összes ilyen esetben a hitel státusza „Fully Paid”, tehát ez nem egy véletlen hiba, hanem adatrögzítési sajátosság.

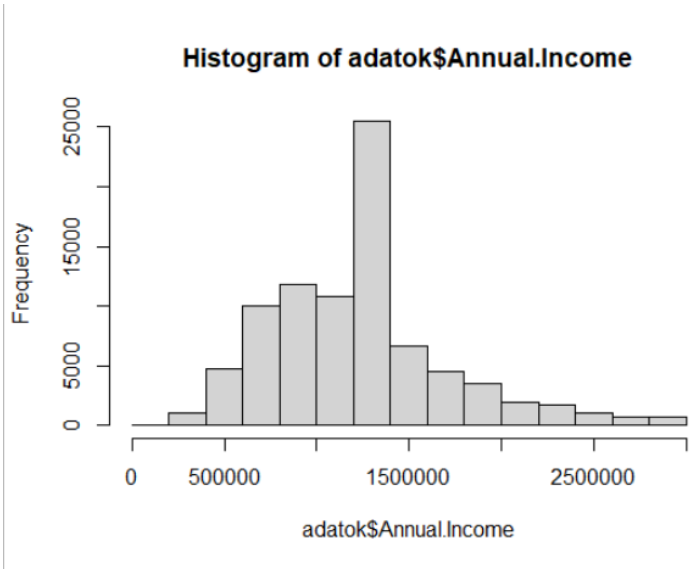
Helyettesítsük ezeket az esteket az esteket az átlaggal (természetesen az átlagot ezen esetek nélkül számítva)! Ezután egy boxplot-tal láthatjuk, hogy néhány ***kiugróan magas érték*** (600.000,- felett) még mindig akad. Ezeket zárjuk ki!

Ezek után a következő hisztogramot kapjuk, ami már egészen jól közelít egy normál eloszláshoz:



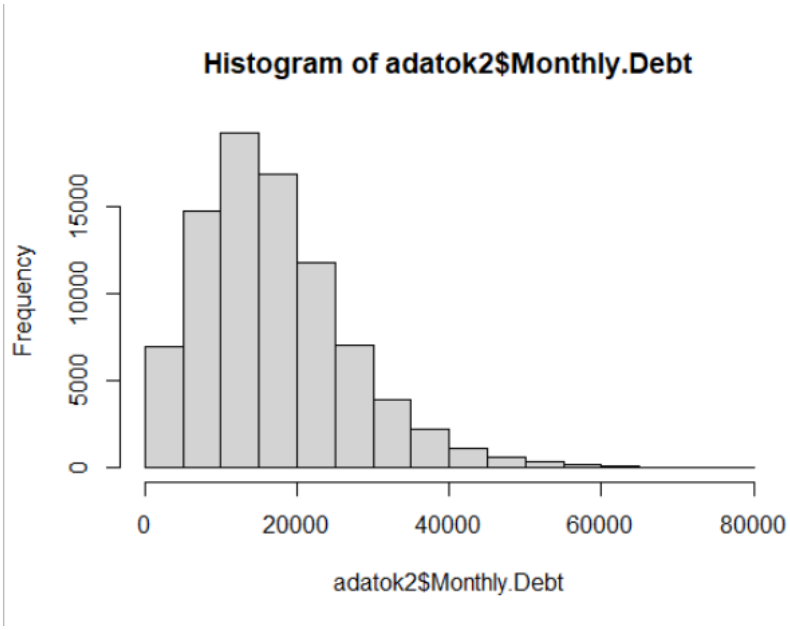
2. sz. ábra: saját forrás (R export)

Az *Annual Income* esetében is nagyon nagy a terjedelem. Megvizsgálva az elosztást, 3.000.000,- feletti összegeket már outilier-nek tekintettem, ezért kizártam az adatokból. Az imputáció miatt az átlagos érték kicsit túlreprezentált, de a vártnál közelebb áll a normál eloszláshoz. A jövedelemmel kapcsolatos mutatók ugyanis rendszerint erősen jobbra elnyúlóak (a nagyszámú kisebb-közepes jövedelműek mellett jellemző a kevés nagyon magas jövedelmű egyed).



3. *sz. ábra: saját forrás (R export)*

A *Monthly Dept* esetében a kigró tételek küszöbét 80.000,- -nél határoztam meg. így egy jobbra elnyúló gyakoriságfüggvényt kapunk, ami szintén reálisnak mondható:



4. *sz. ábra: saját forrás (R export)*

Hasonló logika mellett a *Maxium Open Credit* változónál a 4 Mio feletti értékkel bíró rekordokat zártam ki, és ez az előzőhöz nagyon hasonló gyakoriságot eredményezett.

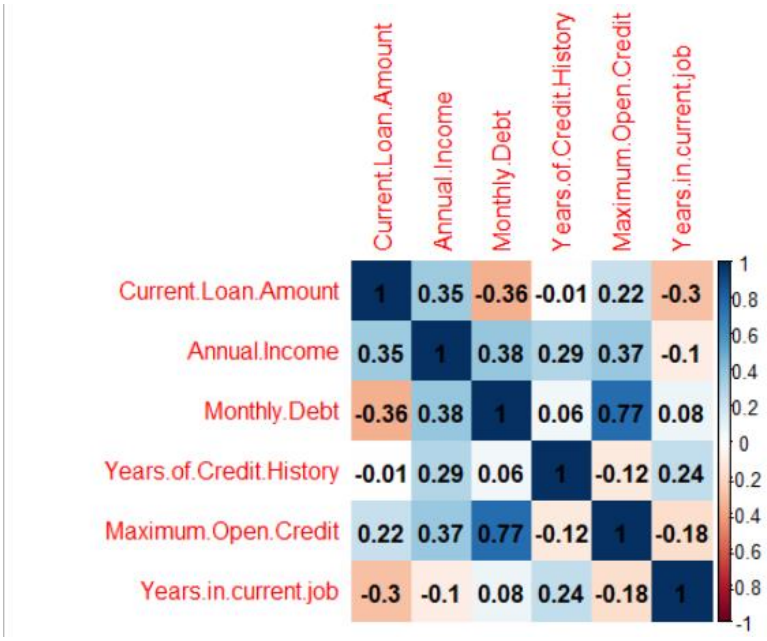
Megvizsgáltam még azt is, hogy nincs-e extrém Debt to Income arány. 13 olyan esetet találtam, melyek meglehetősen különválva a többitől a havi jövedelem több, mint ezerszerese volt a havi adósságtehernek. Ezeket szintén kizártam. A fenti változtatások összesen 15.693 esettel csökkentették adatbázisunkat, tehát még mindig 84.821 sorral rendelkezünk.

A Years in Current Job változó esetén is még célszerűnek mutatkozik egy átalakítás. Ez most egy karakter típusú változó, ám a mögöttes tartalom tulajdonképpen egy ordinális skála (a jelenlegi munkahelyen eltöltött idő években). ezért ezt nominálisra alakítottam át.

Az adatelőkészítés első fázisát ezzel le is zártuk.

9.3 Belső összefüggések, információs tartalom

A nominális változók közti korrelációt az R cor és corplot függvényeivel vizsgáltam. A következő ennek az outputja:



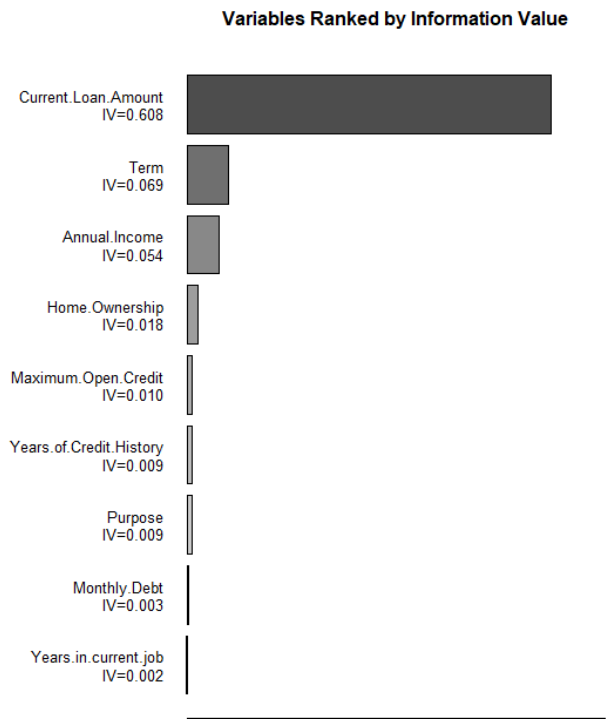
5. sz. ábra: saját forrás (R export)

Mint láthatjuk, *a Monthly Dept és a Maximum Open Credit közt igen erős a korreláció*. Megfontolandó ezért az egyik elhagyása a modellből az esetleges kettős hatás miatt.

A többi esetben csak gyenge-közepesen erős összefüggéseket látunk, ami nem indokol további beavatkozást.

Próbáljuk feltárni, hogy az egyes változók milyen erősen befolyásolják az eredményváltozónk alakulását! Erre a 4. fejezetben ismertetett ***WOE transzformációt illetve az IV-mutatót*** használtam.

Erre az R- ben a Woe-binning csomag nyújt lehetőséget. A változók nominálissá vagy faktorrá alakítása majd a binning elvégzése után a változók IV mutatójáról a következő grafikont rajzolja ki nekünk a program:



6. *sz. ábra: saját forrás (R export)*

Sajnos *a számítás szerint egy kivételével minden változónk információs értéke 0,1 alatti, vagyis gyenge magyarázó erővel rendelkezik!* Sőt, az utolsó négy még a küszöbnek számító

0,02-t sem éri el, tehát magyarázó ereje gyakorlatilag nincs. Ennek oka az lehet, hogy ez az adatbázis nem valós, mesterséges illetve torzított értékeket tartalmaz, azért a belső összefüggései, a magyarázó változók és az eredményváltozók közti kapcsolatok sem érvényesülnek zavartalanul.

Természetesen a WOE tábla ennél részletesebb információkat is ad: az egyes változók „bin”-jeire bontva megmutatja, hogy azokba hány megfigyelés esik, azokban mekkora a célváltozó egyes értékeinek száma, aránya, mekkora az adott bin WOE és IV értéke. Pl. láthatjuk, hogy noha az ügyletek csak 25%-a hosszú lejáratú, a default esetek 35%-a innen kerül ki; A WOE mutatója ezért 41,7, a rövid lejáratú hitelek -16,6-jával szemben. Vagyis a hosszú lejáratú hitelek kockázatosabbak a rövid lejáratúakkal szemben.

9.4 MODELLEK FELÉPÍTÉSE ÉS KIÉRTÉKELÉSE

9.4.1 Döntési fa

Az első modellünk egy döntési fa. Mint azt a már a 1.3. fejezetben is leírtam, a döntési fa tudja kezelni mind a numerikus, mind a kategórikus változókat, és nem „zavarják” kis magyarázó erővel rendelkező változók, mert automatikusan szelektál a változók között. A vágási pontokat az egyes változóknál maga határozza meg, ezért a binning sem feltétlenül szükséges. Ezért a modell paraméterezéséhez már nem szükséges további adatelőkészítő munka.

A döntési fa modellek közül a „C5.0” metodikát választottam. A tanító adatbázisba az összes eset mintegy 70%-át, 64.000 esetet soroltam véletlenszerűen.

A célváltozónk (y) a Loan.Status, magyarázó változók pedig a fent bemutatottak, kivéve az erős korreláció miatt a Maximum .Open.credit-et.[16]

A próbálkozások számát 50-re tettem, aminek eredményeképp a következő modell született meg:

Evaluation on training data (64000 cases):

```
Trial       Decision Tree
-----     ----------------
          Size      Errors

   0        73 14169(22.1%)
   1        11 15876(24.8%)
   2        30 17969(28.1%)
   3        33 20044(31.3%)
   4         6 15465(24.2%)
   5        19 27737(43.3%)
   6        14 17613(27.5%)
   7        19 30870(48.2%)
   8         9 18168(28.4%)
   9        10 21379(33.4%)
  10        12 31808(49.7%)
  11        10 16338(25.5%)
  12         9 35746(55.9%)
  13        15 17192(26.9%)
  14         6 36547(57.1%)
boost        14480(22.6%)   <<


           (a)   (b)    <-classified as
          ----  ----
          1185 13194    (a): class Charged Off
          1286 48335    (b): class Fully Paid


         Attribute usage:
         100.00%        Term
         100.00%        Current.Loan.Amount
          90.75%        Annual.Income
          88.40%        Purpose
          88.40%        Monthly.Debt
          88.40%        Home.Ownership
```

[16] Papírforma szerint az alacsonyabb IV mutató miatt a havi adósságszolgálatot kellett volna kivenni, de aligha találnánk olyan valós credit scoring modellt, amiből ez az adat (vagy a Dept to Income) hiányozna, ezért inkább ennek bent tartása mellett döntöttem.

60.54%	Years.of.Credit.History
52.33%	Years.in.current.job

3. sz táblázat: saját forrás (R export)

A nagyon sok vágás miatt az ágszerkezet kirajzolásának nincs sok értelme, de a fenti összefoglalóból is látjuk, hogy a gép először a futamidő szerint osztotta fel az eseteket, majd a Curent.Loan.Amount különböző értékeinél vágott, majd ezeket az esteket osztotta tovább az Annual.Income majd a Purpose különböző eseteinél – de itt már nem mindig.

Sajnos a változók csekély magyarázó ereje miatt a 15 próbálkozás után az iteráció meg is állt. A további boosting nem volt lehetséges.

A kapott fát ilyenkor még szokásos a *költségmátrix használatával tovább javítani*. Ez azt jelenti, hogy a hibás besorolásokhoz eltérő súlyt adunk (ahogy ezt az 5. fejezetben be is mutattam), amit az algoritmus figyelembe vesz. Ettől azt remélhetnék, hogy a modell szenzitivitása javul (vagyis a FN aránya csökken), ám esetünkben ez sem hozott további javulást.

Ahogy azt már korábban is említettem, lehetséges azt is megvizsgálni, hogy az egyes node-okba mennyi egyed került és ott mekkora a nemfizetés aránya. Vagyis ez a modell is végsősoron eleget tesz annak a szabályozói elvárásnak, ami azt írja elő, hogy az ügyfél nemteljesítési valószínűségét (PD) becsülni kell.

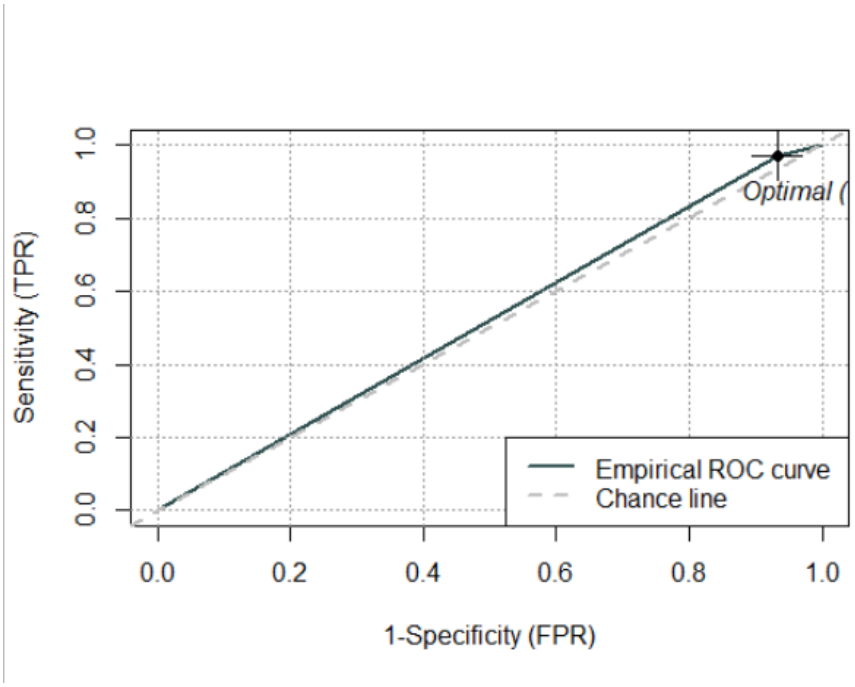
A kiértékelés az 5. fejezetben bemutatott módszerekkel, a modellünknek a teszt adatokon való futtatása után végeztem el.

A konfuziós mátrix alapján a modell pontssága 78,2 %-os, a sensitivity azonban csak 43%-os. Az egyes esetek besorolása:

	Reference	
Prediction	Charged Off	Fully Paid
Charged Off	146	5439
Fully Paid	193	20043

4. sz táblázat: saját forrás (R export)

A felrajzolt ROC-görbe is azt mutatja, hogy a pozitív estek besorolása igen gyenge, alig jobb a véletlenszerű besorolásnál. A görbe alatti terület (AUC): 0,52.



7. sz. ábra: saját forrás (R export)

9.4.2 Logisztikus regresszió

Ahogy a 1.2. fejezetben már említettem, egy logisztikus regressziós modell több megszorítást felételez az adatok tekintetében, mint egy döntési fa. A legfontosabb, hogy nem képes kategorikus, csak numerikus változókat kezelni. Változóink között ilyen a Term, Home Ownership és a Purpose. (jelenlegi munkában eltöltött időt már átalakítottuk.) Ezeket „dummy" változókká kell alakítani, amit az R a „dummyVars" függvénnyel meg is tesz nekünk. Ez a fenti változók minden értékével létrehoz egy újat, és attól függően, hogy az aktuális egyed esetében milyen volt az eredti érték, kerül 1 míg az összes többi esetében

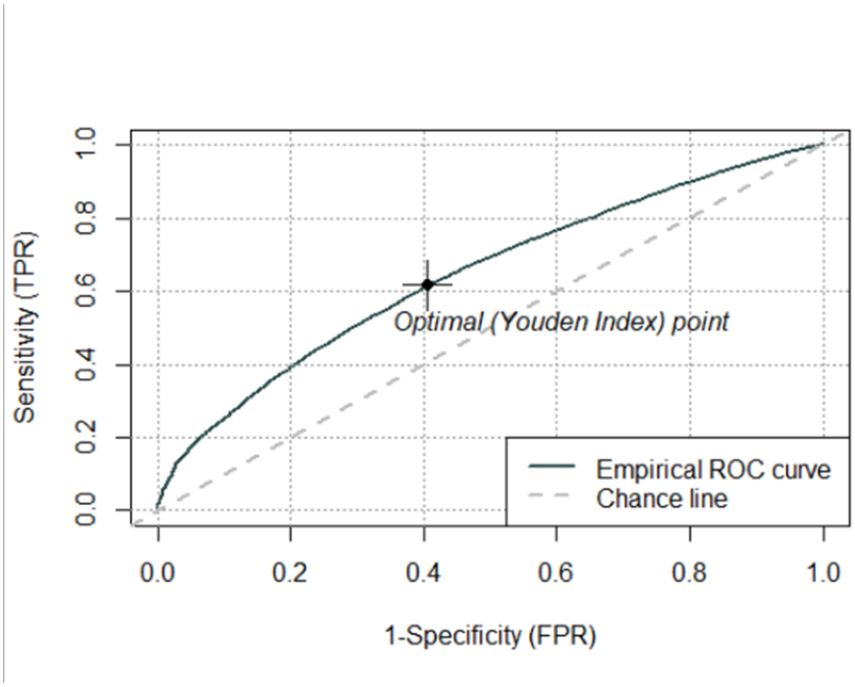
0 az adott új dummy változóba értékként. Esetünkben ez azt jelenti, hogy az eredeti 9 helyett most már 33 változón lett.

A többi előfeltételt (pl. kiugró értékek) már teljesítettük, így felállíthatjuk a modellt. Hasonlóképpen az adatbázist véletlenszerűen felbontjuk tanító és tesz részre, és az R „glm" (a regressziót futtató általános parancs) funkciónál a binominális opció használatával, kijelölve a cél- és a magyarázó változókat, betanítjuk az algoritmust. Ezt következő részben „step" funkcióval tovább pontosítjuk. A modell nem osztályoz, hanem annak az odds-ot adja meg. Az eltárolt modellel a teszt adatbázison kiszámítjuk ezt az értéket, és ezután kell meghatároznunk a cutoff értékét. Ezt az is befolyásolja, hogy mit tartunk a számos mutató közül a leginkább fontosnak (lásd 5. fejezet). Én 0,70-es értéknél találtam meg az optimumot és ezzel a tesztadatokon a következő konfúziós mátrixot kapjuk:

	Reference	
Prediction	Charged Off	Fully Paid
Charged Off	2092	3938
Fully Paid	4852	19939

5. sz táblázat: saját forrás (R export)

Az Accuracy : 0.7148, míg a Sensitivity : 0.30127. Vagyis egy elfogadható pontosság mellett a rossz adósok beazonosítása még mindig nagyon gyenge. A ROC görbét felrajzolva a következőket kapjuk:



8. sz. ábra: saját forrás (R export)

A görbe alatti terület: 0.6455.

Az egyszerű dummy verzióvá átalakítás helyett a *WOE transzformációt* is használhatjuk. Ez azonban ebben az esetben érdemben nem javította az eredményt, így ennek ismertetésétől eltekintettem.

9.4.3 Modellek összehasonlítása

A bevezető során azt mondtam, hgy számunkra az Accuracy mellett a Specificity a legfontosabb mutató. Nos, amennyiben a két alapmodellünket összehasonlítjuk, úgy *az Accuracy vonatkozásában nem volt lényeges különbség, de a specificity a döntési fa esetében jobb volt*. Ráadásul a modell összeállítása, az előfeltételek biztosítása, az adatok előkészítése is ebben az esetben egyszerűbb. Igy – noha nem kaptunk meggyőző eredményt egyik esetben sem - a döntési fa használhatóbbnak tűnik erre az adatbázisra.

10 ÖSSZEFOGLALÁS

A dolgozatomban megpróbáltam betekintést adni arról, hogy a banki hitelezés/kockázatkezelés során milyen szerepet játszanak a minősítő rendszerek, és ezen belül is a statisztikai – adatbányászati modelleken alapuló modellek. Áttekintettem a leggyakrabban használt ilyen metodikákat, kiemelve azok erősségeit és gyengeségeit. Leírtam egy scorecard készítés szokásos folyamatát és a lefontosabb buktatókat, nehézségeket ennek során. Ezután ezen folyamat szerint - de természetesen azt leegyszerűsítve, csak a leglényegesebb lépésekre fokuszálva - a gyakorlatban is próbáltam ezt bemutatni egy teszt-adatbázis segítségével.

A R-ben egy logisztikus regresszesziós modellt és egy döntési fa alapú modellt építettem. A legszükségesebb adatfeldolgozási- átalakítási feladatokat (szintén inkább demonstrációs jelleggel) végrehajtottam az adatbázison. Kiértékelve modelljeinket azt tapasztalhattuk, hogy egy ilyen egyszerűen előállított modell is jobb a véletlenszerű besorolásnál. De az is bebizonyosodott, hogy egy valóban használható, meggyőző diszkriminációs erővel bíró modell felépítéséhez sokkal komolyabb adatelőkészítés és jó adatminőség, valamint kifinomultabb eljárás szükséges.

Az adatok előkészítésnél az előzetes egyváltozós elemzés, a binnelés és a magyarázó érték részletesebb vizsgálata hozhatott volna még valószínűleg magasabb pontosságot, de lehetett volna kisérletezni a leginkább korreláló változók összevonásával is. Az összefüggések feltárásánál pedig a kategórikus változók közti asszociációt lehetett volna elemezni.

További fejlesztési lehetőség más osztályozási metodikák kipróbálása: elsősorban a random forest, a neurális háló és az SVM jöhetne szóba.

Ugyanakkor szeretném hangsúlyozni, a modell építése során elsősorban a folyamat és nehézségek, a leggyakrabban felmerülő problémák és az azokra adható válaszok bemutatása volt, és nem egy, a gyakorlatban is használható modell létrehozása. Úgy vélem, ezt a célt sikerült is elérni.

11 FELHASZNÁLT IRODALOM

Dr. BODON Ferenc (2010): Adatbányászati algoritmusok 2010. http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf letöltve 2023. 04. 08.

BROWN, Iain – MUES, Christophe (2012): An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications 39 (2012) 3446–3453. https://www.sciencedirect.com/science/article/pii/S095741741101342X

FAJSZI Bulcsú – CSER László – FEHÉR Tamás (2010): Üzleti haszon az adatok mélyén. Az adatbányászat mindennapjai; Alinea – IQSYS, 2010

GROTHMANN, Ralph - DISTL, Philipp – KÜHNER, Daniel: A comparison of different machine learning techniques for small ticket credit scoring https://www.academia.edu/37686724/A_comparison_of_different_machine_learning_techniques_for_small_ticket_credit_scoring

HUNYADI László – VITA László (2019): Statisztika közgazdászoknak. KSH, Budapest, 2002. VARGHA András: Többváltozós statisztika dióhéjban: Változó-orientált módszerek. Pólya Kiadó Budapest, 2019.

JÁNOSA András (2015): Adatelemzés IBM SPSS Statistics megoldások alkalmazásával. MKKOK Kft, Budapest, 2015

KISS Ferenc (2003): A credit scoring fejlődése és alkalmazása. 2003 PhD. értekezés. repozitorium.omikk.bme.hu https://repozitorium.omikk.bme.hu/bitstream/handle/10890/258/ertekezes.PDF?sequence=1

LI, Xiao-Lin – ZHONG, Yu (2012): An Overview of Personal Credit Scoring: Techniques and Future Work. International Journal of Intelligence Science, 2012, 2, 181-189 https://www.scirp.org/html/9-1680031_24193.htm

MARKOV, Anton – SELEZNYOVA, Zinaida - LAPSHIN, Victor (2022): Credit scoring methods: Latest trends and points to consider. The Journal of Finance and Data Science7 August 2022. https://reader.elsevier.com/reader/sd/pii/S2405918822000095?token=8AB38E61AC1091830BC967BD46B7D3B73B586D894BCA48E24AFF753015D12FD541787D76BE73300C85A11DF1394AA5B0&originRegion=eu-west-1&originCreation=20230205190716

NYITRAI Tamás (2021): A gépi tanulás módszereinek alkalmazása R-ben Application of machine learning methods in R. Statistikai Szemle, 2021./2 http://unipub.lib.uni-corvinus.hu/7136/1/gepi_tanulas_modszereinek_alkalmazasa_nyt2021.pdf

ORAVECZ Beatrix (2007): Credit Scoring modellek és teljesítményük értékelése. Hitelintézeti Szemle 2007/6.sz.

ORAVECZ Beatrix (2009): Szelekciós torzítás és csökkentése az adósminősítési modelleknél 2009 - phd.lib.uni-corvinus.hu http://phd.lib.uni-corvinus.hu/357/1/oravecz_beatrix.pdf

POLLÁK Zoltán – KOCSIS Ádám (2015): „Minden modell rossz, de némelyikük hasznos” Hitelezési scoring modellek modellezési kockázata. Gazdaság és Pénzügy 2015/1. szám (2. évfolyam)

VARGHA András (2022): Személyorientált többváltozós statisztika: Klasszifikációs módszerek. Pólya Kiadó Budapest, 2022