Biztosítási szerződések díjmentesítési szempontú karakterizációja és klasszifikációja statisztikai adatbányászati módszerek segítségével

MSc szakdolgozat

Takács Kristóf
Biztosítási és pénzügyi matematika MSc

Témavezető:
Vakhal Péter
tudományos munkatárs
Kopint-Tárki Konjunktúrakutató Intézet

Belső konzulens:
Pröhle Tamás
egyetemi tanársegéd
ELTE TTK Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Budapesti Corvinus Egyetem
Közgazdaságtudományi Kar

2018

Tartalomjegyzék

1. Bevezetés **1**

2. A díjmentes leszállítás jelenségének elméleti háttere **3**

3. A díjmentesítési szempontból elemzett életbiztosítási adatbázis általános bemutatása **8**

3.1. Az elemzés során felhasznált adatbázis átfogó áttekintése 8

3.2. Az adatbázis díjmentesítési szempontból érintett elemeinek vizsgálata . 11

4. Az adatbázis díjmentesített szerződéseinek statisztikai adatbányászati karakterizációja és klasszifikációja **13**

4.1. Az alkalmazott statisztikai és adatbányászati módszerek rövid bemutatása 14

4.1.1. Döntési fák és véletlen erdők 14

4.1.2. k-NN (k legközelebbi szomszéd) 16

4.1.3. Naiv Bayes-módszer 18

4.1.4. SVM (support vector machine) 21

4.2. Eredmények 23

4.2.1. Logisztikus bináris regresszió 23

4.2.2. Döntési fák, véletlen erdők 26

4.2.3. k-NN 28

4.2.4. Naiv Bayes-modell 30

4.2.5. SVM 32

5. A szakdolgozat eredményeinek összefoglalása, következtetések **35**

Köszönetnyilvánítás

Ezúton is szeretném megköszönni témavezetőmnek, Vakhal Péternek a számos konzultációt, valamint a sok segítséget és javaslatot, amelyek nélkül ez a szakdolgozat biztosan nem jöhetett volna létre.

Köszönöm továbbá belső konzulensemnek, Pröhle Tamásnak az R programcsomag kezeléséhez nyújtott útmutatásait, illetve a szakdolgozat technikai felépítésére vonatkozó fontos észrevételeit.

Köszönettel tartozom családomnak is, akik a szakdolgozat elkészítésének hónapjai alatt sokféleképpen segítettek nekem, csakúgy, mint egész életem során, továbbá végig biztattak és biztosítottak támogatásukról.

1. fejezet

Bevezetés

Egy életbiztosító állományértékelése során kiemelt jelentőséggel bír az arra vonatkozó feltételezés, hogy az adott élő szerződésállomány tagjai milyen arányban, illetve milyen időbeli kifutással fognak díjmentes leszállítási kérelmet benyújtani. A biztosítótársaságok körében általánosan alkalmazott eljárásmód szerint (elsősorban az adott típusú biztosítással vagy a hozzá hasonló termékekkel kapcsolatos korábbi tapasztalatokra alapozva, amennyiben rendelkezésre állnak felhasználható adatok) az egyes termékekre, illetve termékcsoportokra vonatkozóan előrejelzett díjmentesítési %-ok kerülnek megállapításra, amelyek a tartam eltérő kötvényéveiben jellemzően különböző értékeket vesznek fel.

A rendszeres díjfizetéssel járó biztosítási szerződések díjmentesítése a biztosító jövedelmezőségére nézve negatív következményekkel jár. Ezek közül valószínűleg a legfontosabbnak tekinthető, hogy a díjmentes leszállítás által a biztosító elesik a hátralévő tartamra eső díjbevételtől, amelyen így természetesen nem tud befektetési eredményt elérni. Ugyanakkor a biztosító számára az adott szerződés megszűnéséhez képest általában előnyösebbnek számít a biztosítás díjmentes leszállítása (hiszen általános gyakorlat alapján a társaságok igyekeznek a már meglévő szerződésállomány minél teljesebb körű megtartására, azaz bizonyos speciális esetektől eltekintve az egyes egyéni szintű szerződések állományukban való további kezelésére is), így meglehetősen gyakran alkalmazott eljárás, hogy a szerződő részéről érkező törlési kérelmet a biztosító megpróbálja díjmentes leszállítássá alakítani.

Szakdolgozatomban az egyik magyarországi kompozit biztosító életbiztosítási állományának néhány évvel ezelőtti, egyéni szintű biztosítási szerződési adatait elemezve, modern statisztikai (logisztikus regresszió) és adatbányászati módszerek

segítségével (pl. véletlen erdők, k-NN, support vector machine stb.), az input paraméterek alapján próbáltam meg azonosítani a díjmentesített szerződésekre jellemző, megkülönböztető tulajdonságokat, valamint az ezen esemény bekövetkezésének valószínűségét befolyásoló tényezőket. Egy ilyen típusú elemzés segítséget nyújthat az életbiztosítással foglalkozó cégek számára azáltal, hogy már a szerződés előzetes elbírálása során akár ügyfélszintű predikció adható arra, hogy az adott biztosításon az egyes kötvényévekben milyen valószínűséggel várható díjmentesítés, valamint ennek segítségével pontosíthatóbbá válhatnak a biztosító által belső használatra készített állományértékelések is. Ezenkívül az egyéni szintű díjmentesítési profilok kiértékelésének lehetőségén túl az alkalmazott módszerek megfelelő eszközt jelenthetnek a termék-, illetve termékcsoport szintű díjmentes leszállítási százalékok pontosabb kalkulációjához, ezáltal akár a biztosítások árazásának megfelelőbbé tételéhez is.

A szakdolgozat szerkezeti felépítését áttekintve a diplomamunka középpontjában álló jelenség elméleti hátterének bemutatását az elemzési céllal átadott adatbázis átfogó jellegű vizsgálata követi, amely magában foglalja az adatok általános elemzése mellett a díjmentesítési szempontból kiemelt fontosságú adatbázis-elemek értékelését is. Ezután kerül sor az alkalmazott módszerek rövid bemutatására, majd a felhasználásukkal kapott eredmények ismertetésére, végül az utolsó fejezet egy összefoglalást tartalmaz a szakdolgozatban elért eredményekhez kapcsolódóan.

2. fejezet

A díjmentes leszállítás jelenségének elméleti háttere

A díjmentesítés fogalma meglehetősen régóta (kb. a XIX. század végétől), azzal szorosan összefonódva jelen van a biztosítási szektorban, így nem meglepő, hogy kimondottan széles körű az ezen jelenség hátterét vizsgáló szakirodalom. Szakdolgozatom jelen fejezetében a díjmentes leszállítás folyamatához kapcsolódó fontosabb könyvrészleteket és tanulmányokat mutatom be és értékelem őket, kitérve a legjelentősebbekre a hazai és a nemzetközi források közül egyaránt.

Banyár József a díjmentes leszállítást mint a (vegyes biztosítási) díjtartalékhoz kapcsolódó maradékjogok[1] egyikét tárgyalja. A szerző az ügyfél biztosítási tartalékhoz fűződő jogait a biztosító szemszögéből kategorizálva ismerteti, az ügyfél likviditási nehézségének súlyosságát középpontba állítva:

- A díjmentes leszállítást a szerződő tartós, a nemfizetés kockázatát is tartalmazó likviditási nehézségének egy lehetséges megoldásaként mutatja be.
- A kötvénykölcsönt a díjmentesítéssel párhuzamba állítva elemzi, amely egy az előzőhöz hasonló típusú, viszont átmeneti jellegű probléma elhárítására szolgáló eszköz lehet a biztosító és a kötvénytulajdonos közötti kapcsolatban.
- A visszavásárlási opció pedig a kutató szerint tipikusan egy olyan tartós és súlyos negatív anyagi változás bekövetkezését feltételezi a szerződőnél,

[1]Maradékjog: egy biztosító azon díjtartalékra vonatkozó elszámolási kötelezettsége, amely egy biztosítási szerződés valamilyen okból történő felbomlása során keletkezik.

amelyben nemcsak a díjfizetést kell megszakítania, hanem az adott pillanatban szükségessé válik számára a felhalmozott díjtartalék felhasználása is.

(A szerző megemlíti, hogy bár a kötvénykölcsön nem tekinthető valódi maradékjognak, könyve logikai felépítése megkívánja, hogy az adott fejezetben foglalkozzon ezen biztosításhoz kapcsolódó jelenséggel; illetve kiemeli, hogy annak ellenére, hogy jogi szempontból a díjmentes leszállítás már maradékjognak tekinthető, véleménye szerint az valójában csupán egy „standardizált szerződésátdolgozási lehetőség".) [3]

A szerző megfogalmazása szerint a díjmentes leszállítás folyamata során a biztosító a szerződés adott pillanatban meglévő díjtartalékát egy az eredetivel megegyező típusú, viszont nem rendszeres, hanem egyszeri díjas biztosítás díjaként kezeli. Az így keletkező új szerződés tartama megegyezik a díjmentesítés érvényesítése és az eredeti biztosítás tartamának utolsó napja közötti időtartammal, az új biztosítási összeg pedig a szerződő aktuális életkora alapján kerül meghatározásra. A díjmentes leszállítás fogalmának definiálása után Banyár József rátér a maradékjogok korlátozásainak okaira és jellemző előfordulásukra az egyes fontosabb biztosítástípusokat tekintve. Megállapítja, hogy (elsősorban az antiszelekció veszélye miatt) a haláleseti életbiztosítások esetén tipikusan minden maradékjog kizárásra kerül a biztosítók részéről, míg a tiszta elérési biztosításoknál (és így a járadékbiztosításoknál is) jellemzően a visszavásárlás és a díjmentesítés lehetőségét nem szokták engedélyezni. Olyan konstrukciókban viszont, amelyekben az alapvetően különálló módon értékesített haláleseti és elérési biztosítást egy csomagban, azaz tulajdonképpen vegyes biztosításként kínálják, viszonylag általános a maradékjogok engedélyezése is azzal a kikötéssel, hogy az adott csomag egyik részét alkotó biztosításon történő maradékjog-érvényesítés automatikusan együtt jár ugyanezen jognak a csomag többi részén való életbe lépésével. (Ezzel a feltétellel kiküszöbölhető az antiszelekció esélyének drasztikus megnövekedése, amely különben reális és valószínű veszélyként jelentkezne.) [2]

Tompa Krisztina cikkében [15] a díjmentes leszállítást mint egy a hagyományos életbiztosításokhoz köthető opciót kezeli, és ennek a lehetőségnek az egyes szerződésekre vonatkozó értékét próbálja statisztikai modellezés, valamint különböző szimulációk segítségével meghatározni. A szerző az életbiztosítási szerződések implicit opcióit két fő csoportba sorolja, amelyek közül a díjmentesítést a garanciákkal

szemben a szerződő jogai között tünteti fel, hasonlóan az újrakezdési opcióhoz (resumption option), a visszavásárláshoz, a dinamikus díjkiigazításhoz és a garantált járadékhoz. A szerző által alkalmazott megközelítés a díjmentes leszállításra adott definícióban a díjfizetés tartam alatt történő felfüggesztésének lehetőségét emeli ki, amely ebben az esetben a biztosítás fedezetének megmaradása mellett tud megvalósulni, hiszen ekkor az ügyfél „a díjmentes leszállítás időpontjáig felhalmozott tartalékból mint egyszeri díjból egy alacsonyabb biztosítási összegű szerződést kap".

Ezenkívül megemlítésre kerül a díjmentesítéshez szorosan kapcsolódó újrakezdési (visszatérési) opció is, amely lehetővé teszi a szerződő számára, hogy a díjmentes leszállítás végrehajtását követően a díjfizetés újrakezdését kezdeményezze, ezzel egyidejűleg a biztosítási összeg jelentős megemelését is elvégezve, amely folyamat eredményeképpen tipikusan az eredeti volument megközelítő biztosítási összeg alakulhat ki. A visszavásárlási opció bemutatása során a szerző párhuzamot von a díjmentes leszállítással arra alapozva, hogy mindkét esetben megszűnik a díjfizetés, ugyanakkor véleménye szerint mind a biztosító, mind fedezeti szempontból az ügyfél számára egyaránt előnyösebbnek tekinthető a díjmentesítés (a meglévő állomány megtartása, illetve a szerződés érvényben maradása miatt). [15]

A Tompa Krisztina cikkében biztosítási opció árazási célra használt szerződés egy klasszikus vegyes életbiztosítás nyereségszámla alapú többlethozam-visszatérítéssel, amely modellben a díjmentesítés (pontosabban azon év, amelyben bekövetkezik) először mint paraméter jelenik meg: a szerződő minden kötvényév elején eldöntheti, hogy kíván-e díjmentesíteni vagy sem. Néhány bevezető vizsgálatot követően viszont a szerző modelljében áttér az egyes kötvényévekhez meghatározott díjmentesítési százalékokat rendelő profil használatára, ezzel egyúttal az életbiztosítóknál a gyakorlatban alkalmazott módszerekhez közelítve az elemzett cash-flow szcenáriók struktúráinak peremfeltételeit. Az opció értékének meghatározása a többlethozam-visszatérítési százalékok felhasználásával történt: a szerző azt vizsgálta, hogy változatlan pénzáramot feltételezve a szerződő mint racionális döntéshozó mekkora visszajuttatásról hajlandó lemondani azért cserébe, hogy rendelkezzen a díjmentesítés lehetőségével.

A szimulációs futtatások során elemzésre kerültek a díjmentesítési opcióra a szerződés egyéb paraméterei által gyakorolt hatások is: a cikk kitér többek között a hozamgörbe, a belépési kor és a biztosítás típusa által determinált érzékenységek vizsgálatára. A tanulmány alkotója összefoglalásként megállapítja, hogy az előzetes

várakozásoknak megfelelően a hozamgörbe felfelé tolódása az opció értékére negatív, a lefelé mozdulás viszont pozitív hatással rendelkezett (a hozamokban történő 100 bázispontos változás kb. 50%-os csökkenést, illetve közel 100%-os növekedést eredményezett az egyes esetekben). A szimulációs futtatások outputjának belépési kor szerinti analízise megmutatta, hogy a szerződés megkötésénél megfigyelhető biztosítotti életkor növekedése ceteris paribus a díjmentes leszállítási lehetőség értékének folyamatos csökkenésével jár együtt. A biztosítás típusa szerinti vizsgálatok alapján pedig kijelenthető, hogy egy elérési biztosítás esetén a díjmentesítés opciója jelentősen magasabb értéket képvisel, mint egy kockázati életbiztosításnál. Ugyanakkor fontos megállapítás, hogy a cikkben bemutatott kutatás alapján tipikusnak tekinthető szerződésparaméterek és piaci peremfeltételek mellett a díjmentesítési opció értéke nem nevezhető számottevőnek. [15]

Nadine Gatzert tanulmányában [8] a díjmentes leszállítást elsősorban mint a szerződő ún. „nemteljesítési opciójának" (nonforfeiture option) egyik lehetséges megvalósulási formáját tárgyalja a visszavásárlás, valamint (elsősorban whole life típusú biztosítások esetén) egy kiterjesztett tartamú biztosítás létrehozása mellett. A szerző megközelítésében a nemteljesítési opció biztosítja, hogy a szerződés korai (a biztosított halálát megelőző) lezárása esetén a vonatkozó jogszabályok előírásainak megfelelően a szerződő tulajdonát képező díjtartalék nem veszhet el, hanem annak méltányos része (pl. visszavásárlás esetén a visszavásárlási büntetéssel csökkentett összeg) a szerződőt kell hogy megillesse. A nemteljesítési opcióból adódó kötelezettségek rendezésével összefüggő kifizetések összegének természetesen bármely megoldási forma esetén megfelelő módon összevethetőnek kell lennie a szerződés adott pillanatban aktuális aktuáriusi értékével.

A szerző cikkében egy táblázat formájában igyekszik egy szándékai szerint teljes körű listát adni az életbiztosításokhoz kapcsolódó implicit opciókról. Megközelítése szerint ezen csoportosítás tizenöt nagyobb kategóriát tartalmaz, amelyek közül a díjmentes leszállítás a „díjfizetésre vonatkozó opciók" közé tartozik (az újrakezdési, a dinamikus díjkiigazítási és a rugalmas díjfizetési lehetőségek mellett). (Érdekességként megemlíthető, hogy az „osztalékopciók", azaz a többlethozam felhasználására vonatkozó lehetséges alternatívák között megtalálható a „díjmentesített szerződés hozzáadása" is.) A tanulmány szerzője megállapítja továbbá, hogy a díjmentes leszállítás valójában egy forward szerződésre vonatkozó put opcióként is kezelhető, amelynek tárgya, hogy az opció lehívása után a szerződő

vállalja, hogy $K = 0$ kötési árfolyamnak megfelelő éves díjú kötvényeket fog vásárolni. Ezzel a put opcióval megvalósíthatóvá válik, hogy hatása kizárólag arra korlátozódjon, hogy az eredeti szerződést egy új, díjfizetés nélküli és csökkentett kifizetésű biztosítássá transzformálja. [8]

3. fejezet

A díjmentesítési szempontból elemzett életbiztosítási adatbázis általános bemutatása

Szakdolgozatom ezen fejezete az egyik hazai életbiztosító által elemzési céllal részemre átadott adattáblák áttekintő jellegű bemutatását tartalmazza. Az áttekintés mélységét az alkalmazott statisztikai módszerekben felhasznált adatokkal összhangban próbáltam meghatározni, amely eljárások részletes ismertetése, illetve az így kapott eredmények a 4. fejezetben találhatók. Az első alfejezet teljes mértékben általános célú bemutatását követően a diplomamunka témájának megfelelően áttérek a vizsgált időpontban megfigyelhető díjmentesen leszállított állomány fontosabb paramétereinek értelmezésére.[1]

3.1. Az elemzés során felhasznált adatbázis átfogó áttekintése

Az átadott adatok érzékenysége miatt, illetve titoktartási okokból meg nem nevezett biztosító által rendelkezésemre bocsátott adatbázisok az állomány várható jövőbeni alakulásának modellezésére felhasznált, a vállalat által kezelt hagyományos

[1] A szakdolgozat nyilvánosan hozzáférhető változatában az átadott adatok titkosságának a biztosító kérésére történő megőrzése céljából az adatbázis elemeinek változónként egy-egy rögzített nagyságú konstans értékkel való módosítássa történt meg.

életbiztosítási termékekre vonatkozó, szerződés szintű adatokat tartalmaztak, amelyek az egyik széles körben ismert modellező program inputját alkották. Az ezen körbe tartozó adatokhoz a 2011. év végi állapotnak megfelelően tudtam hozzáférni, a változók áttekintését az 1. táblázat tartalmazza.

Változónév	Jelentés	Skála
pol_comm_y	szerződéskötés éve	intervallum
pol_comm_m	szerződéskötés hónapja	ordinális
gender	szerződő neme	nominális
birth_year	szerződő születési éve	intervallum
birth_month	szerződő születési hónapja	ordinális
age_modifier	a valódi és a kalkuláció során figyelembe vett életkor eltérése (év)	intervallum
pol_term	szerződés tartama (év)	intervallum
index_pc	indexálás mértéke (%)	intervallum
rnp_start	nyereségtartalék (Ft)	intervallum
comm_sum	2012-2031 között kifizetni tervezett jutalék összege (Ft)	intervallum
si_act_ver	biztosítási összeg (Ft)	intervallum
prem_act_y	aktuális éves biztosítási díj (Ft)	intervallum
prem_freq	díjfizetési gyakoriság	nominális
sum_prem_mp	összes befizetett biztosítási díj (Ft)	intervallum
years_since_start	a szerződés kezdete óta eltelt idő (év)	intervallum
age_start	a szerződő életkora a szerződés megkötésekor (év)	intervallum
pup_dummy	2011-ig díjmentesített szerződésekre vonatkozó bináris változó	nominális
pup_term	díjmentesítés óta eltelt idő (év)	intervallum

1. táblázat: A felhasznált változók elnevezése, magyarázata és típusa.

A 2011-es tradicionális biztosítási termékek állománya 39 665 db modellpontból állt, amely szerződések átlagos tartama kb. 15,5 év volt. Egy hagyományos termékkel rendelkező tipikus ügyfél átlagosan 1973-ben született, azaz a keresztmetszeti vizsgálat időpontjában megközelítőleg 37-38 éves volt. A nemek közötti megoszlást elemezve megfigyelhető, hogy a hagyományos termékek körében a férfi-női arány meglehetősen kiegyenlítetten alakult (50,45% - 49,55%). Az egyes szerződésekre az akkori időponttól számított 20 éven belül kifizetendő átlagos jutalék összege kb. 9000 Ft volt, viszont ezen középérték mögött kimondottan aránytalan megoszlás húzódott meg, amelyre pl. a nagymértékű relatív szórás is utal.

Változónév	Átlag	Szórás
pol_comm_y	2008,79	4,77
pol_comm_m	7,67	3,42
gender	1,50	0,50
birth_year	1973,83	11,60
birth_month	7,41	3,43
age_modifier	1,08	0,72
pol_term	15,32	13,23
index_pc	2,10	3,96
rnp_start	96688,48	119449,25
comm_sum	8976,66	29088,45
si_act_ver	752427,71	1551652,29
prem_act_y	33717,34	224256,39
prem_freq	7,65	5,14
sum_prem_mp	52096,83	137600,67
years_since_start	2,65	4,78
age_start	34,64	12,06
pup_dummy	0,08	0,27
pup_term	0,57	2,28

2. táblázat: A 2011. év végi hagyományos szerződésállomány fontosabb statisztikai mutatói.

A biztosítási összegeket vizsgálva látható, hogy az adott biztosítónál egy hagyományos életbiztosítási szerződés átlagosan 750 000 Ft-os nagyságú paraméterrel rendelkezett ezen változót vizsgálva. (Érdemes megemlíteni, hogy a 2011. év végi állomány legmagasabb biztosítási összege elérte a 70 millió Ft-ot.) A szerződésekre vetített átlagos éves díj valamivel 34 000 Ft alatt alakult (a díjmentes szerződések nélkül ezen érték kevéssel meghaladta a 36 000 Ft-ot). Az ügyfelek közel 60%-a a biztosítási díj havi rendszerességgel történő befizetését preferálta, 15%-uk negyedévente, illetve 4%-uk félévente rendezte díját, ugyanakkor több mint 10%-uk az évente egy összegben történő rendezést választotta. A vizsgált szerződések átlagosan kb. 2,7 évvel a 2011-es mérlegfordulónap előtt kerültek megkötésre (a szerződők átlagéletkora a megkötés időpontjában 34,5 év volt), illetve ekkor még 12,7 év volt hátra lejáratukig.

3.2. Az adatbázis díjmentesítési szempontból érintett elemeinek vizsgálata

A biztosító 2011-es tradicionális szerződésállományának több mint 8%-ánál, vagyis 3191 db szerződésnél kezdeményeztek az ügyfelek korábban díjmentes leszállítást. A díjmentesített szerződések átlagos tartama jelentősen meghaladta az összállományét, hiszen ezen paraméter 21 évet megközelítő értéke a teljes adatbázisban megfigyelhető adatnál több mint 35%-kal magasabb volt. A díjmentesített szerződéssel rendelkező ügyfelek születési évének átlaga 1972-re esett, ezáltal kb. 39 éves átlagéletkort eredményezve ezen csoporton belül, így ebből a szempontból nem volt lényeges eltérés tapasztalható a kiindulási adatbázishoz viszonyítva.

Változónév	Átlag	Szórás
pol_comm_y	2003,06	3,15
pol_comm_m	7,33	3,35
gender	1,48	0,50
birth_year	1972,49	10,17
birth_month	7,41	3,39
age_modifier	1,00	0,27
pol_term	20,84	3,01
index_pc	8,64	5,66
rnp_start	167279,81	167049,98
comm_sum	5028,76	1164,70
si_act_ver	558659,85	586097,27
prem_act_y	5000,00	0,00
prem_freq	8,89	4,48
sum_prem_mp	304669,68	407462,99
years_since_start	8,41	3,14
age_start	30,25	9,85
pup_dummy	1,00	0,00
pup_term	7,04	4,40

3. táblázat: A 2011 végéig díjmentesített szerződésállomány fontosabb statisztikai mutatói.

A nemek közötti megoszlást megfigyelve alapvetően az összesített adatokhoz hasonlóan nagyságrendileg többé-kevésbé azonos méretű csoportok voltak azonosíthatók, ugyanakkor érdemes megjegyezni, hogy a díjmentes leszállítások körén belül a női szerződők voltak enyhe többségben (52%). A díjmentesítés folyamatának sajátosságait tekintve Szintén a díjmentes leszállítás sajátosságait tekintve nem nevezhető meglepőnek, hogy az ebbe a csoportba tartozó szerződésekre érzékelhetően alacsonyabb átlagos biztosítási összeg volt jellemző: ez a kb. 550 000 Ft-os érték a teljes állomány megfelelő paraméterének 70%-át sem érte el. (Jellemző adat, hogy a legmagasabb díjmentes biztosítási összeg kb. 10,4 millió Ft-ot ért el, szemben a teljes állomány 70 millió Ft-os maximumával.)

A díjmentesített szerződéseknél kiemelten érdekes lehet megvizsgálni, hogy az egyes szerződők összesen mennyi biztosítási díjat fizettek be az adott időpontig (amely összeg ebben az esetben természetesen megegyezik a díjmentesítés végrehajtásának pillanatáig befizetett pénzmennyiséggel). Ezen változót elemezve látható, hogy ezen ügyfelek átlagosan 305 000 Ft biztosítási díj befizetése után kezdeményezték a díjmentesítést, ugyanakkor meglehetősen nagy változékonyság volt jellemző ezen paramétert tekintve, hiszen a mutató szórása majdnem elérte az átlagérték 1,35-szeresét, egyebek mellett annak következtében, hogy az ezen kategóriába tartozó szerződések több mint 5%-a 1 millió Ft-ot meghaladó befizetéssel rendelkezett (az egy díjmentesített szerződésre befizetett legnagyobb volumenű díj összege meghaladta a 7,5 millió Ft-ot).

A díjmentesen leszállított szerződések átlagosan 8,5 évvel a 2011. év végi zárás előtt kerültek megkötésre, amely időtartamból kb. 7 évet töltöttek díjmentesített állapotban, lejáratukig pedig ekkor még kb. 12,5 év volt még hátra. A szerződések aláírásakor a későbbi díjmentesítők átlagéletkora 30 év volt, amely érték jelentősen alacsonyabb volt a teljes állományra jellemző hasonló paraméternél (átlagosan 35 éves korban történő szerződéskötés).

4. fejezet

Az adatbázis díjmentesített szerződéseinek statisztikai adatbányászati karakterizációja és klasszifikációja

Diplomamunkám ezen fejezetében bemutatom a különböző matematikai és statisztikai alapú eljárások eredményeit, amelyek segítségével a rendelkezésre álló adatok felhasználásával megpróbáltam a lehető legnagyobb pontossággal karakterizálni az adatbázisban szereplő azon szerződéseket, amelyek 2011. év végéig díjmentes leszállításon estek át.[1]

Az egyes alfejezetek felépítését tekintve először azon alkalmazott módszerek rövid bemutatása szerepel, amelyek nem képezték a szak törzsanyagának részét, majd részletesen ismertetem az ezen eljárások, illetve a bináris logisztikus regresszió által a szakdolgozatban elemzett adatbázis inputként történő felhasználásával szolgáltatott kimeneti eredményeket.

[1] Mivel a bemutatandó klasszifikációs módszerek működése az adatok titkossága megőrzése céljából elemenként és változónként végrehajtott konstans eltolásra érzéketlenek, így ezen fejezet a dolgozat nyilvános változatában is az eredeti, nem torzított értékeket tartalmazó adatbázis felhasználásával kapott eredményeket mutatja be.

4.1. Az alkalmazott statisztikai és adatbányászati módszerek rövid bemutatása

4.1.1. Döntési fák és véletlen erdők

A statisztikai adatbányászat területéhez tartozó módszerek közül a legszélesebb körben elterjedtek közé tartoznak a döntési fák és a véletlen erdők, amelyek mind regresszió, mind klasszifikáció típusú feladatok megoldásában alkalmazhatók. Mivel szakdolgozatomban egy alapvetően klasszifikációs jellegű problémát próbálok különböző módszerekkel megoldani, az eljárás bemutatása során is egy osztályozási probléma megoldásán keresztül szemléltetem annak működését.

Egy klasszifikációs feladat általános végrehajtási sémája a következőképpen fogalmazható meg: az adatpontok bizonyos bemeneti változói felhasználásával az adott modell egy előre meghatározott értékekkel (osztálycímkékkel) rendelkező kategorikus változó valamely értékét adja vissza outputként. Az input változók elnevezése: *prediktor*ok, a kapott kimeneti érték pedig az egyes adatpontokra vonatkozó *predikció*.

A döntési fa modell működésének alapötlete viszonylag természetesnek nevezhető: minden lépésben rekurzív módon úgy próbáljuk több (általában kettő) kisebb részre osztani az aktuálisan vizsgált adathalmazt, hogy az így keletkező új adathalmazok valamilyen értelemben a lehető legközelebb kerüljenek egy olyan állapothoz, amelyben lehetőleg már csak azonos osztályba tartozó adatpontok vannak jelen egy-egy minél homogénebb output halmazban. Szemléletesen legtöbbször egy gyökérrel rendelkező bináris faként szokás ábrázolni a modellt, amelynek minden csúcsa megfelel egy szeparációs feltételnek: a prediktorai alapján az adott feltételt teljesítő adatpontok az adott csúcs valamely gyerekébe kerülnek, míg a feltételnek nem megfelelő egyedek a másik gyerekcsúcsba. [6]

A módszer fontos kérdései közé tartozik, hogy az egyes csúcsokban melyik prediktor alapján történjen meg a szeparáció, illetve ezen belül is az adott input változóra vonatkozóan pontosan milyen feltétel kerüljön megállapításra. Ezen modellparaméterek meghatározására klasszifikációs feladatoknál tipikusan a Gini-indexet szokták mint heterogenitási mutatót alkalmazni: minden változó és minden felosztás szerint kiszámítják a $G = 1 - \sum_{i=1}^{c} p(i|t)^2$ értékek súlyozott átlagát a

keletkező gyerekcsúcsokra, ahol $p(i|t)$ az i-edik címke előfordulási gyakoriságát jelöli a t csúcsban lévő elemek között.

A fenti súlyozott átlag minimalizálása ekvivalens az adott csúcsban lévő elemek optimális megbontásával az egy lépésben elérhető maximális lehetséges homogenitást figyelembe véve: az $(1 - G)$ mutató annál nagyobb értéket vesz fel, minél „sikeresebb" (homogénebb gyermekcsúcsokat eredményező) az aktuális kettéosztás. Kategorikus változók esetén az ezen input paraméter értékeit figyelembe véve előállítható összes lehetséges megbontás kipróbálható, numerikus változóknál viszont ez nem kivitelezhető, így általában az adott numerikus változó által felvett értékek bizonyos finomságú diszkretizálása valósul meg (pl. minden kvartilisnél vagy decilisnél történő elvágás tesztelése). [4]

A döntési fa modell számos előnye mellett (könnyű értelmezhetőség, intuitivitás stb.) jelentős hátrányának nevezhető, hogy hajlamos a túlillesztésre: bár az ismert, a modell létrehozására felhasznált adatbázis elemeit akár meglehetősen nagy pontossággal képes pontosan besorolni, ugyanakkor ezen jelenség kialakulásában tipikusan közrejátszik, hogy a kialakuló döntési fa struktúrája az adathalmaz egyedi sajátosságait is leképezi, amely feltételek egy adatbázison kívüli, ismeretlen egyed besorolása során feltehetően irrelevánsak, így az osztályozás pontosságát rontják. A túlillesztés jelensége által gyakorolt hatások csökkentése érdekében két fő módszert szoktak alkalmazni: egyrészt a fa felépítése után az egyes, jellemzően kis létszámú levelekben kialakuló homogén osztályok „metszéssel" (pruning) megszüntethetők, azaz így inhomogén levelek keletkeznek, amelyekhez tartozó címke a többségi elv alapján kerül meghatározásra, ezáltal az input adathalmaz specifikus jellemzőinek figyelembevétele feltehetően kisebb mértékben történik meg, mint az eredeti modellben.

A túlillesztés kiküszöbölésére gyakran alkalmazott másik fő módszer a véletlen erdők használata. Egy véletlen erdő több döntési fa halmazából áll, ezen különböző fák ugyanazon adatpontra más-más predikciót adnak outputként, így az erdő által adott végső klasszifikációs besorolás ezen különböző predikciók összessége alapján történik meg (általában a többségi elvet figyelembe véve). A modell nevében is szereplő véletlenséget (hiszen a döntési fa felépítése egy determinisztikus algoritmus elvégézését jelenti) többnyire kétféle módon lehet biztosítani:

- A *bagging* módszerben az egyes fák felépítése során különböző adathalmazok kerülnek felhasználásra, amelyek véletlen visszatevéses mintavételi módszerrel keletkeznek az input adatokból.

- A *változók variálása* módszer során a különböző fák ugyanazon input adathalmazzal dolgoznak, viszont az elvágás során csak az eredeti változók egy véletlen részhalmazát veszik figyelembe, amelynek mérete nem haladhatja meg az összes változó számának négyzetgyökét.

Számos gyakorlati kísérlet, illetve elméleti eredmények is igazolják, hogy a véletlen erdők a döntési fákhoz képest kevésbé hajlamosak a túlillesztésre, illetve jellemzően nagyon pontos predikcióra képesek, hátrányuk viszont többek között, hogy kevésbé intuitívak és vizualizálhatók. [5]

4.1.2. k-NN (k legközelebbi szomszéd)

A statisztikai adatbányászat területén a véletlen erdők mellett régóta használt és népszerű klasszifikációs (és regressziós) eljárásként említhető többek között az úgynevezett k-NN (k legközelebbi szomszéd) módszer. Az eljárás mögött lévő alapgondolat szemléletesen a következőképpen fogalmazható meg: a modell felépítése során azzal a feltételezéssel élünk, hogy (egy klasszifikációs problémát tekintve) az aktuálisan osztályozni kívánt egyedhez a numerikus prediktorok többdimenziós terében „közel lévő" adatpontok várakozásaink szerint nemcsak ezen bemeneti változók, hanem a célváltozó értékeit figyelve is hasonlóak lesznek egymáshoz.

A fentieket precízebben megközelítve kijelenthető, hogy a k-NN módszer feladatának célja egy új megfigyelés besorolása egy y kategorikus változó valamely osztályába úgy, hogy a kiinduló adatbázis elemeire ismertek y és az $x_1, \ldots, x_p$ prediktorok értékei is, az aktuálisan vizsgált egyedre vonatkozóan viszont csupán az x_i prediktorok állnak rendelkezésre. A besorolás a prediktorok által meghatározott térben az új egyedhez legközelebb eső k db adatpont y szerinti címkéjének leggyakoribb elemével egyezik meg, ahol k egy előre rögzített, pozitív egész értékű paraméter. (Fontos feltétel, hogy minden x_i prediktornak olyan numerikus változónak kell lennie, amelyekre az általuk felvehető értékek közötti távolság értelmezhető: az intervallum vagy skála szinten mért változók esetén ez természetesen kivitelezhető, viszont a modell felépítésébe kategorikus változók is bevonhatók, amennyiben a kategóriacímkék közötti „távolságok" megfelelően definiálhatók.) [1]

A módszer által az O új egyedre outputként adott osztályozás az alábbiak szerint formalizálható:

$$\mathrm{cat}(O) = \arg\max_{c \in C} \ |\{O_i \in N_k(O) : \mathrm{cat}(O_i) = c\}|.$$

A fenti képletben $N_k(O) = \{O_1, \ldots, O_k\}$ az O egyedhez az x_i prediktorok által generált térben értelmezett d távolság szerint legközelebb eső k pont halmazát jelöli, $\mathrm{cat}(O_i)$ pedig az O_i megfigyelés y szerinti osztályozását, ahol az y lehetséges értékeit a C halmaz tartalmazza. (Az $N_k(O)$ halmaz O középpontú gömbök sugarának folytonos átmenetű növelésével kerül előállításra.)

A k-NN módszer kiemelt fontosságú előkészítő lépései közé tartozik, hogy szükséges az y kategorikus célváltozót tekintve irreleváns prediktorok eltávolítása a bemeneti adatok közül, mivel egyrészt ezek a változók a vizsgálandó többdimenziós tér dimenzióját feleslegesen megnövelik, másrészt mivel nem szignifikánsak a célváltozó szempontjából, így értékeik figyelembevétele a predikció végrehajtásakor szintén kerülendő, alkalmazásuk a besorolás pontosságát jellemzően rontja. A fentiek mellett a skálázási hatás kiküszöbölése érdekében a változók sztenderdizálását is el kell végezni, ezáltal a keletkező prediktortér koordinátatengelyei azonos beosztásúak lesznek, amely így lehetővé teszi az ezen térben megfigyelhető távolságok korrekt kiszámítását és a prediktorok azonos súllyal történő kezelését.

Az előzőekben bemutatott eljárás több technikai jellegű kérdést is felvet: például előfordulhat, hogy az $N_k(O)$ halmaz előállítása során kapott gömbben k-nál több adatpont található, viszont a gömb sugarának bármilyen kis mértékű csökkentésével kevesebb mint k pont kerülne az új, kisebb gömbbe (azaz $N_k(O)$ felszínén szükségszerűen több pont is megfigyelhető, amelyek mindegyikének hozzávétele a szomszédsági halmazhoz annak méretét túlságosan megnöveli). Több lehetséges megoldás is elképzelhető ezen probléma kezelésére: pl. az ehhez hasonló esetekben megengedhető, hogy az összes (k-nál több) $N_k(O)$-beli pont figyelembevételre kerüljön a végső besorolás meghozatalánál; de az is elképzelhető, hogy ha minden esetben k db pont alapján szeretnénk döntést hozni, akkor a gömb felszínén lévő pontok közül (véletlenszerűen vagy valamilyen egyéb szempont szerint optimalizálva) csak megfelelő számút veszünk bele a döntési halmazba. Szintén technikai problémának tekinthető, ha az $N_k(O)$-beli elemeket vizsgálva több olyan osztálycímke is megfigyelhető, amelynek előfordulási gyakorisága ezen halmazon belül maximális.

A probléma jellemző megoldása, hogy az ilyen jellegű esetekben a predikciót ezen leggyakoribb kategóriák közül történő véletlenszerű kiválasztás adja meg.

A k-NN eljárás alkalmazását, illetve a kialakuló modellek besorolási pontosságát jelentős mértékben befolyásolja, hogy milyen k paraméter mellett kerül futtatásra a fenti algoritmus. Fontos speciális eset a $k = 1$ választás (legközelebbi szomszéd módszer), amely viszonylag intuitív módon adódik, viszont a gyakorlati tapasztalatok szerint az így kialakuló modellek tipikusan kevéssé robusztusak, illetve túlillesztésre is hajlamosak; $k = n$ esetén ugyanakkor minden új megfigyelésre az input adatbázis y-ra vonatkozó módusza lesz az adott predikció, amely gyakorlati szempontból szintén nem használható. Az optimális k értékének meghatározása például a validációs adatbázison történő, különböző nagyságú paraméterekre történő teszteléssel valósítható meg, amely eljárással kiválasztható az a k^*, amelyre a téves besorolások aránya minimális, így a tesztadatbázison már ezen k^* beállításával futtatható az algoritmus. [12]

4.1.3. Naiv Bayes-módszer

A gyakorlati tapasztalatok szerint a bonyolultabb módszerekhez hasonlóan jó eredményeket szolgáltató adatbányászati klasszifikációs eljárás az úgynevezett naiv Bayes-módszer (amely a véletlen erdőkkel és a k-NN-nel szemben kizárólag osztályozási feladatok megoldására használható, regressziós problémák kezelésére nem alkalmas). Az eljárás alapelve a Bayes-tételre épül: ha A és C olyan események, amelyekre $P(A) > 0, P(C) > 0$ teljesül, akkor

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

A fenti képletben a $P(A|C)$ mennyiség szokásos elnevezése *a priori*, $P(C|A)$-é pedig *a posteriori* valószínűség. Ezek a megnevezések onnan származnak, hogy ha A-t egy megfigyelhető eseménynek tekintjük, akkor $P(A|C)$-t egy C előzetes feltevésből, hipotézisből származó (a priori) valószínűségként kezelhetjük, amely a fenti képlet szerinti kapcsolatban áll azzal az előrejelzésre, predikcióra vonatkozó (a posteriori) valószínűséggel, amelyből megállapíthatjuk, mekkora a C hipotézis helyességének valószínűsége, amennyiben az A esemény következett be. [11]

A Bayes-tétel előzőekben ismertetett értelmezési alternatívája viszonylag kézenfekvő módszert biztosít egy klasszifikációs feladat megoldására: az input

attribútumokat valószínűségi változóként kezelve a tesztadatbázis x_i prediktorai alapján kiszámítható a priori valószínűségek segítségével minden y_j kategóriacímkére meghatározhatók az a posteriori valószínűségek, amelyek közül a legnagyobb értékhez tartozó besorolás kerül elfogadásra egy új modellpontra vonatkozóan. Formalizálva a fentieket a következő összefüggés adódik:

$$P(Y = y_j \mid \mathbf{x} = (x_1, \ldots, x_p)) = \frac{P(\mathbf{x} = (x_1, \ldots, x_p) \mid Y = y_j)P(Y = y_j)}{P(\mathbf{x} = (x_1, \ldots, x_p))},$$

amely képletben az $\mathbf{x} = (x_1, \ldots, x_p)$ vektor a prediktorok által az adott adatpontra felvett értékeket tartalmazza, Y pedig a célváltozónak megfelelő valószínűségi változót jelöli.

Az input változók függetlenségének feltételezése mellett a nevezőben szereplő, a prediktorok együttes eloszlására vonatkozó valószínűség egyszerűbb formában írható:

$$P(Y = y_j \mid \mathbf{x} = (x_1, \ldots, x_p)) = \frac{\prod_{i=1}^{p} P(X_i = x_i \mid Y = y_j)P(Y = y_j)}{P(\mathbf{x} = (x_1, \ldots, x_p))},$$

ahol X_i az i-edik prediktornak megfeleltetett valószínűségi változót jelöli. Az egyenlőség jobb oldalán szereplő mennyiség nevezője nem függ Y-tól, így a keresett a posteriori valószínűség maximalizálásának feladata ekvivalens a jobb oldali tört számlálójának maximalizálásával, amelynek összes tényezője becsülhető a teszt adatbázis elemeinek segítségével: a $P(Y = y_j)$ valószínűségek az y-ra vonatkozó relatív gyakoriságokkal, a $P(X_i = x_i \mid Y = y_j)$ értékek pedig a prediktorokra jellemző, az Y célváltozó szerinti egyes kategóriákon belüli relatív gyakoriságokkal kerülnek közelítésre ezen eljárásban, amennyiben az X_i prediktor nem numerikus változó.

A naiv Bayes-féle megközelítés lényeges különbségének nevezhető az előző alfejezetben ismertetett k-NN eljáráshoz képest, hogy a modell által alkalmazott x_i prediktorok tetszőleges skálán mérhetőek lehetnek (tehát a nominális és az ordinális változók sincsenek kizárva), ugyanakkor meg kell felelniük bizonyos feltételeknek: a kategorikus változóknak az Y célváltozó minden osztályában függetleneknek kell lenniük, illetve ezenkívül a numerikus változók Y kategóriáiban normális eloszlást kell hogy kövessenek. (A klasszifikálási metódus elnevezésében szereplő „naiv” jelző a prediktorokra vonatkozó ezen függetlenségi feltevésre utal.)

A normalitási feltételezés miatt a numerikus prediktorokra elegendő a feltételezett $X_i|_{Y=y_j} \sim \mathrm{N}(\mu_{ij}, \sigma_{ij})$ eloszlású valószínűségi változó paramétereit megbecsülni, ezáltal a $P(X_i = x_i \mid Y = y_j)$ feltételes valószínűség az előző változó sűrűségfüggvényének megfelelő helyen felvett értékével közelíthető:

$$P(X_i = x_i|Y = y_j) \approx \frac{1}{\sqrt{2\pi}\sigma_{ij}} \, e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

A gyakorlati tapasztalatok szerint a naiv Bayes-módszer elfogadható hatékonyságú működéséhez valóban szükséges a fenti heurisztikus jellegű levezetésben megfogalmazott feltételek (megfelelően kis hibaértékkel történő) teljesülése. Emiatt célszerű egymástól független prediktorokat választani: az összefüggő magyarázó változók közül a kevésbé fontosnak feltételezettek elhagyhatók, vagy valamilyen dimenziócsökkentési eljárással (pl. főkomponenselemzés) egy új, összevont változóvá alakíthatók. A numerikus változók esetén ezenkívül érdemes a normalitási kritérium teljesítésére is törekedni, ennek érdekében a prediktorokon különböző transzformációk is elvégezhetők (pl. logaritmizálás).

Kategorikus magyarázó változóknál megfigyelhető probléma, hogy ha az adott adatbázisra a $P(X_i = x_i \mid Y = y_j)$ feltételes valószínűségek közelítésére alkalmazott relatív gyakoriságok közül valamelyik 0 értéket ad, akkor a maximalizálni kívánt érték automatikusan szintén 0-ra kerül beállításra ezen prediktorok és y_j megfigyelése esetén, amely jelenség a besorolás pontosságát ronthatja. A módszer ezen negatív tulajdonsága kiküszöbölése céljából a $P(X_i = x_i \mid Y = y_j) \approx \frac{N_{ij}}{N_j}$ approximáció helyett (ahol N_{ij} az X_i prediktor x_i értékkel való megegyezésének előfordulásának számát jelöli az y_j kategórián belül, N_j pedig az ezen célváltozó kategóriába tartozó összes elem számát) gyakran a Laplace-féle $P(X_i = x_i \mid Y = y_j) \approx \frac{N_{ij}+1}{N_j+c}$ becslést alkalmazzák (ahol c a lehetséges y_j kategóriák számának felel meg). Ezenkívül előfordul az m-esztimátorok használata is: ekkor $P(X_i = x_i \mid Y = y_j) \approx \frac{N_{ij}+mp}{N_j+m}$, ahol p a becsülni kívánt a priori feltételes valószínűségre adott előzetes közelítésnek felel meg, m pedig egy paraméter, amelyet viszonylag kis értékként érdemes választani (az $m \in [1, 2]$ választás nevezhető általánosnak). [9]

4.1.4. SVM (support vector machine)

A machine learning módszerek közé tartozó SVM (support vector machine; 'támaszvektor-gép') eljárás meglehetősen sokoldalú célra felhasználható, a tapasztalatok szerint nagy pontosságú eredményeket előállító metódusnak nevezhető: klasszifikációs, regressziós, sőt klaszterezési feladatok megoldására is alkalmas. (A szakdolgozat témájához illeszkedve ezen alfejezetben csak osztályozási problémákra vonatkozóan mutatom be az SVM működését.) Kiemelt alkalmazási területei közé tartozik többek között a különböző nemstrukturált adathalmazok feldolgozása (pl. szövegek, képek, hangok, kézírás azonosítása, besorolása). A módszer fontos jellemzőjének nevezhető, hogy kizárólag numerikus prediktorok kezelésére képes, ugyanakkor előnynek tekinthető, hogy nem véletlent használó algoritmus, hanem végrehajtása a kezdeti paraméterek beállítása után teljes mértékben determinisztikus módon történik. [13]

A support vector machine modell legegyszerűbb változata az ún. *lineáris SVM*: ezen módszer célja az $y \in \{-1, 1\}$ bináris kategorikus változó értékei szerinti két csoportba tartozó modellpontoknak az $x_1, \ldots, x_p$ prediktorok p-dimenziós terében egy olyan $(p-1)$ dimenziós hipersíkkal való tökéletes elválasztása, amelyre teljesül, hogy az egyes csoportoktól mért távolsága a lehető legnagyobb. Az ezen elválasztó hipersíkkal párhuzamos, a két csoport „legszélső" elemeire illeszkedő hipersíkok elnevezése: „támaszvektorok" (support vector), amelyekről a módszer az elnevezését is kapta. (Pl. a $p = 2$ esetben olyan egyeneseket kell keresni, amelyek a síkot úgy osztják két részre, hogy az y szerinti csoportokat besorolási hiba nélkül szeparálják, miközben távolságuk maximális.)

Az elválasztó hipersík egyenletére valamilyen $\mathbf{w}$ és b mellett $\mathbf{wx} = b$-nek kell teljesülnie, így a hipersík egyik oldalán lévő párhuzamos vektorokra $\mathbf{wx} - b > 0$, a másik oldalon lévőkre pedig $\mathbf{wx} - b < 0$ áll fenn. Megfelelő konstans szorzó alkalmazásával a $\mathbf{wx} - b$ érték az első támaszvektorra 1-et, a másikra pedig -1-et vesz fel, továbbá belátható, hogy a két támaszvektor távolsága $\frac{2}{||\mathbf{w}||}$-vel egyezik meg, így a lineáris SVM megoldása a következő szélsőérték-feladat optimumának megtalálásával ekvivalens:

$$\min_{\mathbf{w},b} ||\mathbf{w}||$$
$$y_i(\mathbf{wx}_i - b) \geq 1$$
$$(i = 1, \ldots, n; y_i \in \{-1, 1\})$$

Lényeges kritérium, hogy a fenti gondolatmenet csak akkor érvényes, ha az y célváltozó szerinti két csoport lineárisan szeparálható, azaz létezik olyan $(p-1)$ dimenziós hipersík, amely a prediktortér ezen kategória szerinti felosztását tökéletesen végre tudja hajtani. Amennyiben ez nem megoldható, akkor célszerű lehet egy olyan $\phi : \mathbb{R}^p \to \mathbb{R}^q$ nemlineáris leképezés keresése, amely az adatpontok eredeti p dimenziós terét úgy transzformálja át $\mathbb{R}^q$-ba, hogy a modellpontok képe ebben a magasabb dimenziójú térben y szerint már lineárisan szeparálható. A *nemlineáris SVM* módszer ezen alapötlete a Cover-tételre épül, amely szerint két lineárisan nem szétválasztható csoportra magasabb dimenziójú, nemlineáris transzformációt alkalmazva nagy valószínűséggel az adatpontok olyan képhalmaza áll elő, amely már lineárisan szeparálható. [7]

A nemlineáris SVM feladat az előzőek alapján az alábbi probléma megoldását jelenti:

$$\begin{gathered} \min_{\mathbf{w},b} ||\mathbf{w}|| \\ y_i(\mathbf{w}\phi(\mathbf{x}_i) - b) \geq 1 \\ (i = 1, \ldots, n; y_i \in \{-1, 1\}) \end{gathered}$$

A ϕ transzformáló függvény meghatározását általában nem közvetlenül, hanem a hozzá tartozó ún. κ magfüggvényen (kernel) keresztül érdemes elvégezni. A $\kappa : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ függvény az új térbe áttranszformált adatpontok közötti skaláris szorzatot adja meg: $\kappa(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$. A ϕ és a κ függvények kölcsönönösen meghatározzák egymást, ugyanakkor a magfüggvény használatának jelentős előnye, hogy lehetővé teszi, hogy a nemlineáris SVM feladat optimalizálása kizárólag skaláris szorzatok kiszámításával valósuljon meg.

A gyakorlatban általánosan elterjedt kernelfüggvények közé tartozik például a Gauss-féle radiális bázisfüggvény (RBF): $\kappa(\mathbf{x}, \mathbf{y}) = e^{-\gamma||\mathbf{x}-\mathbf{y}||^2}$, ahol $\gamma > 0$ paraméter. Ezenkívül szintén gyakran alkalmazott magfüggvénynek számít a polinomiális kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \gamma(\mathbf{x} \cdot \mathbf{y} + c)^d$ ($\gamma > 0, c \geq 0, d$ pozitív egész paraméterek), illetve a szigmoid kernel is: $\kappa(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x} \cdot \mathbf{y} + c)$ ($\gamma > 0, c \geq 0$ paraméterek). (A $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ lineáris kernel választásával visszakapható az eredeti lineáris SVM feladat.)

A nemlineáris SVM módszer problémájának nevezhető, hogy csak abban az esetben alkalmazható, ha az $\mathbb{R}^q$-ba áttranszformált adatpontok már lineárisan szeparálhatók. Egy ezen feltételt teljesítő ϕ (vagy az ezzel ekvivalens κ kernel)

megtalálása azonban tipikusan nem könnyen megoldható probléma, hiszen a két csoport tökéletes szétválaszthatósága meglehetősen erős feltételt jelent. Ezen a szigorú kritériumon enyhít a *modern SVM* által alkalmazott megközelítés, amelyet kifejlesztése óta a gyakorlatban az SVM feladatok megoldására jellemzően alkalmazni szoktak.

A modern SVM módszer megengedi, hogy a két csoportot elválasztó hipersík ne biztosítson teljes mértékű szeparációt. Egy adott H hipersík mellett minden $\mathbf{x}_i$ adatpont esetén kiszámításra kerül az ε_i hibaérték: $\varepsilon_i = 0$, ha $\mathbf{x}_i$ a H által meghatározott, az y_i kategóriához tartozó támaszvektor megfelelő oldalán található; $\varepsilon_i \in (0, 1]$, ha $\mathbf{x}_i$ az adott oldali támaszvektor és az elválasztó H hipersík között helyezkedik el (tehát a H szerinti klasszifikáció ebben az esetben $\mathbf{x}_i$-re helyes besorolást ad); és $\varepsilon_i > 1$, ha a H által adott osztályozás nem megfelelő $\mathbf{x}_i$-re (az ε_i hiba nagysága az y_i kategória támaszvektorától mért távolsággal arányosan nő).

A modern SVM feladat a fentiek szerint a következőképpen formalizálható:

$$\min_{\mathbf{w},b,\varepsilon} ||\mathbf{w}||$$
$$y_i(\mathbf{w}\phi(\mathbf{x}_i) - b) \geq 1 - \varepsilon_i$$
$$\varepsilon_i \geq 0$$
$$\sum_{i=1}^{n} \varepsilon_i \leq C$$
$$(i = 1, \ldots, n; y_i \in \{-1, 1\})$$

Ebben a felírásban C egy előre meghatározott paraméter, amely az adott modern SVM modell kategorizálása által tartalmazott, az input elemek nem megfelelő osztályozásából származó hibák összegére adott felső korlátot jelöli. [7]

4.2. Eredmények

4.2.1. Logisztikus bináris regresszió

A bináris klasszifikáció céljára alkalmazható egyik legalapvetőbb módszer a logisztikus regresszió, amely eljárással a 2011 végéig díjmentesített szerződésállomány azonosítását próbáltam végrehajtani (célváltozó: *pup_dummy*). A backward Wald metódus többszöri iterációját követően (minden lépésben az aktuális modell

legkevésbé szignifikáns változóját manuálisan eltávolítva az inputból) az 5. táblázatban szereplő output adódik, amely modellben szereplő minden magyarázó változó együtthatója szignifikánsan eltér a 0-tól, így a modell által szolgáltatott egyéb eredmények is értelmezhetőek.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 9[c]	pol_term	,042	,002	311,988	1	,000	1,043
	si_act_ver	,000	,000	86,051	1	,000	1,000
	mp_start	,000	,000	164,054	1	,000	1,000
	sum_prem_mp	,000	,000	317,487	1	,000	1,000
	years_since_start	,360	,008	1813,355	1	,000	1,433
	Constant	-7,163	,121	3492,024	1	,000	,001

5. táblázat: A backward Wald eljárással kapott modell.

Az output illeszkedési mutatóiból látható, hogy a kapott modell közepesen jól illeszkedik az input adatokra: a Nagelkerke-féle R^2 több mint 75 %-os értéket ért el, valamint a log-likelihood mutató nagysága is szignifikánsan magas értékű volt (6756,48). A modell eredményeként kialakuló leverage értékek minden adatpont esetében 0,1 alatt maradtak, a Cook-távolság pedig csupán egy esetben haladta meg az 1-es küszöbértéket, azaz az adattábla méretét figyelembe véve számottevő mértékben nem voltak azonosíthatóak a regresszió kialakítását egyedileg jelentősen befolyásoló adatpontok. [14]

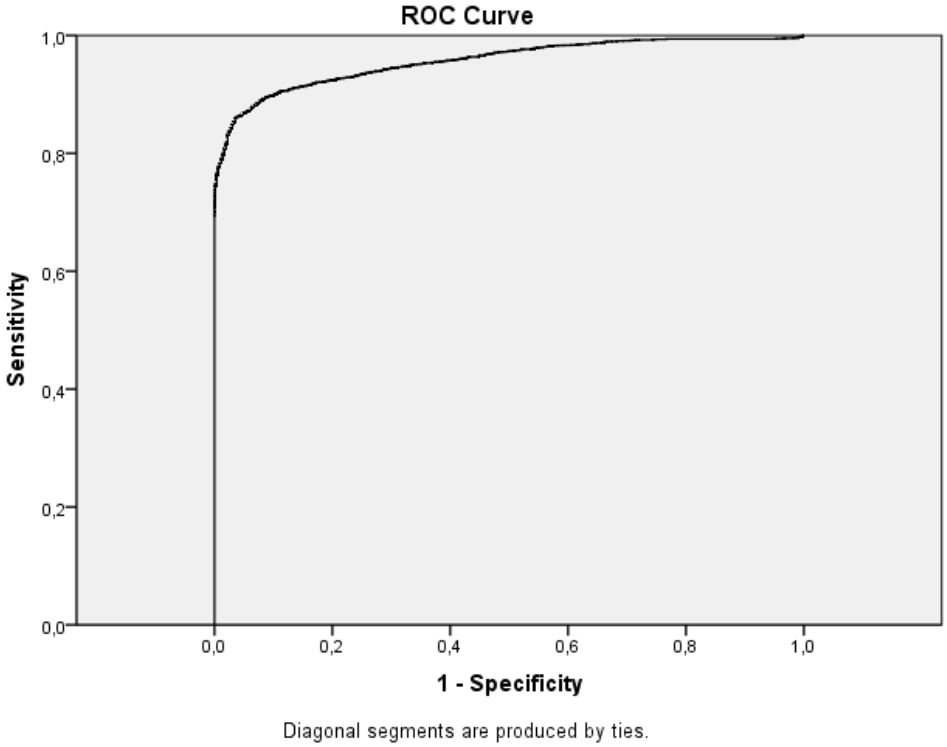
Classification Table[a]

			Predicted		
			pup_dummy		
	Observed		0	1	Percentage Correct
Step 9	pup_dummy	0	35047	1427	96,1
		1	440	2751	86,2
	Overall Percentage				95,3

a. The cut value is ,080

6. táblázat: A backward Wald eljárással kapott modell klasszifikációs táblája.

Figyelembe véve az adatbázis nagy aranyú kiegyenlítetlenségét a célváltozó értékeinek szempontjából, szükségessé vált a cut-off paraméter alapbeállításának jelentős mértékű módosítása. A predikciós valószínűségeken alapuló szegmentációs változó értékét a mintabeli gyakorisághoz illeszkedő módon 0,08-ként választva adódó klasszifikációs tábla eredményeiből kiolvasható, hogy a modell összességében 95%-os találati aránnyal tudta besorolni a magyarázó változók alapján a szerződéseket a két csoportba; a parciális találati arányokból látható, hogy míg a díjfizető szerződéseket kimondottan magas, 96% fölötti arányban sikerült helyesen kategorizálni, ezzel szemben a díjmentesített szerződéseket csupán 86%-os pontossággal osztályozta megfelelően. A regressziós módszer szerint kiszámított besorolási valószínűségek alapján felrajzolható ROC görbe alatti terület megközelítette a 0,96-ot, ebből következően ez a mérőszám a modell átlagosnál jobb illeszkedését prognosztizálta (ld. 1. ábra). [10]



1. ábra: A backward Wald eljárással kapott modell ROC görbéje.

A magyarázó változók különböző együtthatóit és szignifikanciaszintjeit tartalmazó táblázat adataiból megállapítható, hogy a logisztikus regresszió a bemeneti

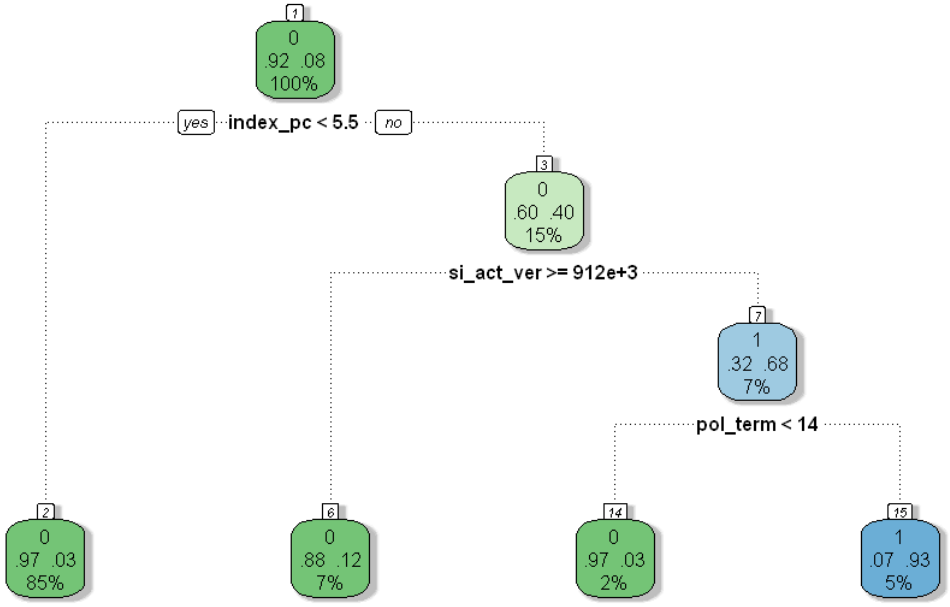
magyarázó változók közül végül csak ötöt (tartam, a szerződés kezdete óta eltelt idő, biztosítási összeg, nyereségtartalék, a vizsgálat időpontjáig befizetett összes biztosítási díj) használt fel a szerződések besorolására. A logit értékekből az exponenciális függvény transzformációja által kapható valószínűség-szorzókat tartalmazó Exp(B) oszlopot vizsgálva levonható a következtetés, amely szerint a modell kimeneti eredményei alapján a szerződés kezdete óta eltelt idő növekedése átlagosan a díjmentes leszállítás bekövetkezésének valószínűségét nagymértékben növeli, illetve kisebb mértékben, de szintén ilyen irányú változást eredményez a hosszabb szerződéstartam is.

4.2.2. Döntési fák, véletlen erdők

Az R programcsomag *rpart* könyvtárát használva a 2. ábrán látható döntési fa keletkezik, amennyiben a *pup_dummy* változó értékére szeretnénk predikciót adni egy legfeljebb 3 mélységű döntési fa segítségével. Megállapítható, hogy a kapott modell a szerződés megkötése során rögzített indexálási százalékra, a biztosítási összegre, illetve a szerződés tartamára vonatkozó változókat használja fel az adathalmazok optimális szétvágásához. Ezen döntési fa alapján kizárólag az 5,5%-os indexálásúnál magasabb, 912 ezer Ft-nél alacsonyabb biztosítási összeggel és a legalább 14 éves tartammal rendelkező szerződések lennének díjmentesként klasszifikálhatók, amely modell viszont egy meglehetősen alacsony arányú (kb. 60%-os) pontossághoz vezet a díjmentesen leszállított szerződések sikeres besorolását vizsgálva.

A létrehozandó modell mélységére vonatkozó feltételt elhagyva egy 6 mélységű fa keletkezett, amelynek felépítésében már többek között a nyereségtartalék összegére vonatkozó változó is szerepelt. Ugyanakkor még ez a modell is csupán 72%-os pontossággal volt képes helyesen besorolni a díjmentesített szerződéseket, így célszerűnek tűnt egy véletlen erdő elkészítése is ezen klasszifikációs probléma vizsgálata céljából.

Az R *randomforest* csomagját használva különböző paraméterbeállítások mellett több véletlen erdőt is előállítottam, amelyek mindegyike 500 döntési fát tartalmazott; a fák felépítése során minden esetben három véletlenszerűen kiválasztott változó került felhasználásra. Az egyes erdők klasszifikációs tábláját megvizsgálva látható volt, hogy még a viszonylag soknak számító 100 egyedet tartalmazó csúcsot is további szétvágás nélkül kezelő modell ($nodesize = 100$) is jelentősen magasabb találati arányt (kb.

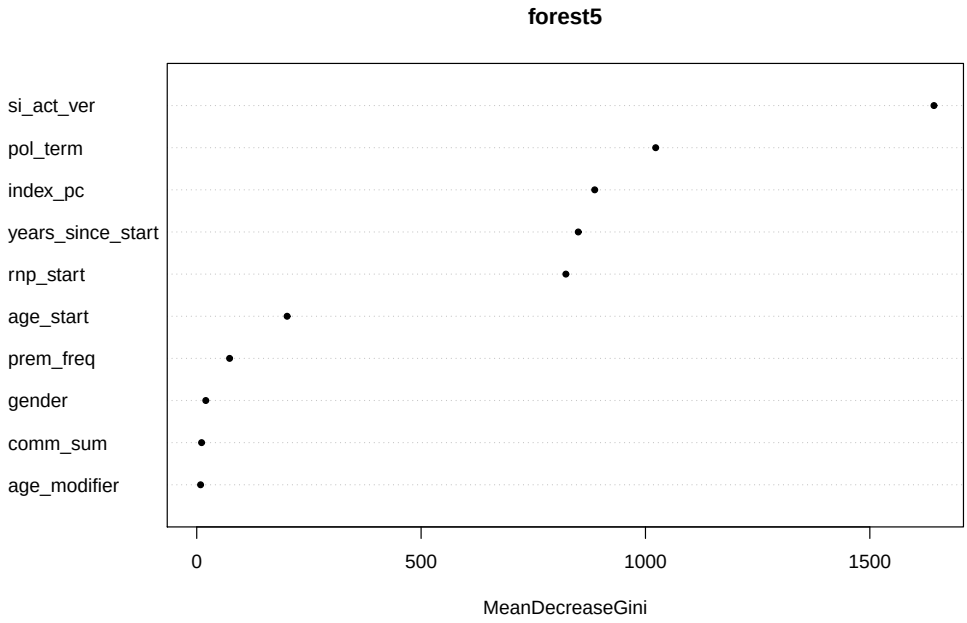


2. ábra: Az elemzett adatbázis segítségével felépített 3 mélységű döntési fa.

79%) ért el a díjmentes szerződések besorolásánál, mint az optimális döntési fa. Ezen találati pontosság a megengedett maximális csúcsméret csökkentésével tovább nőtt: a paraméter egy-egy csúcsban 50, 25, illetve 5 szerződést megengedő beállítása 83, 85, illetve 87%-os találati pontosságot eredményezett. (Bár természetesen - különösen az utolsó esetben - felmerülhet a kérdés, hogy a véletlen erdők túlillesztéssel kapcsolatban tapasztalható kedvező tulajdonságai az adatbázis struktúrája tulajdonságainak ilyen mértékű felhasználása esetén is érvényben maradnak-e.)

A véletlen erdők jelentős hátrányának tekinthető, hogy felépítésükből adódóan viszonylag nehéz megállapítani, hogy a modell kialakításában az egyes változók mekkora jelentőséggel rendelkeztek. Ennek a problémának a megoldására biztosít egy korlátozott lehetőséget a varImpPlot parancs, amelynek outputjaként kirajzolódó diagram a felhasznált változókat csökkenő sorrendbe rendezi aszerint, hogy az erdőt alkotó fák felépítése során az adott változó értékei mentén történő szétvágások átlagosan mekkora csökkenést okoztak a súlyozott Gini-mutatók kiszámításában.

A 3. ábra varImpPlot diagramjáról egyértelműen leolvasható, hogy a legkisebb csúcsméretet megengedő véletlen erdőben az egyes szerződések besorolásának elvégzése során átlagosan a biztosítási összeg mentén történő szétvágások eredményezték a lokálisan legnagyobb mértékű homogenitási változást az adott csúcs

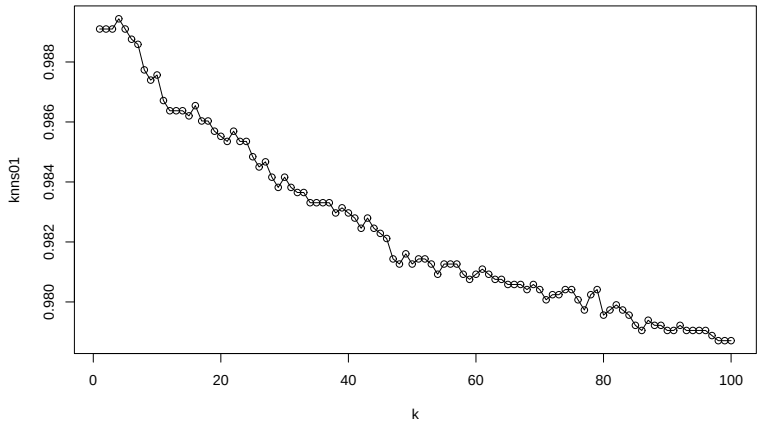


3. ábra: A változók „fontosságát” ábrázoló diagram ($nodesize = 5$).

gyermekeit vizsgálva. Ebből a szempontból fontosabb változónak volt tekinthető ezenkívül a szerződés tartama, az indexálási faktor, a szerződés megkötése óta eltelt idő és a nyereségtartalék aktuális értéke, ezzel szemben a fizetési gyakoriság vagy a szerződő neme kevésbé jól szeparáló változónak bizonyult a döntési fák felépítésénél.

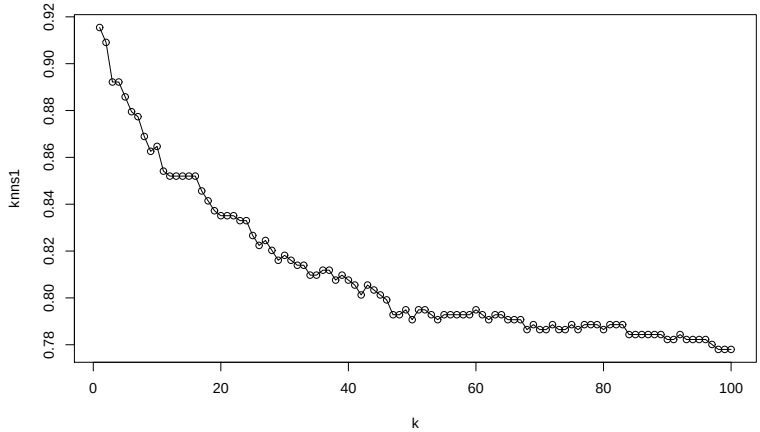
4.2.3. k-NN

A k-NN módszer alapvető sajátosságai szükségessé tették, hogy elvégzésre kerüljön az input adatbázis két részre osztása: véletlenszám-generálás segítségével az összesen megközelítőleg 40000 adatpont kb. 85%-a randomizált módon a tanító adathalmazba került, a fennmaradó 15% pedig a tesztelési célra alkalmazott adatbázist alkotta. Az R programcsalád *class* könyvtárát használva a k paraméter értékének változtatásával ($k \in \{1, 2, \ldots, 100\}$) 100 különböző k-NN modellt építettem fel a 4.2.1. alfejezet végleges logisztikus regressziós modelljében szereplő szignifikáns változók $(pol\_term, si\_act\_ver, rnp\_start, sum\_prem\_mp, years\_since\_start)$ sztenderdizált értékeit mint prediktorokat kezelve, célváltozóként továbbra is a díjmentesítés bekövetkeztét jelölő pup_dummy változót alkalmazva.



4. ábra: A tesztadatbázis elemeinek helyes besorolási aránya az elkészített k-NN modellekben különböző k értékek mellett.

A teljes tesztadatbázisra vonatkozó találati pontosság kiemelkedően magas értékeket vett fel: mint ahogy az a 4. ábráról is leolvasható, az összes tesztelt k értékre $98 - 99\%$ közötti besorolási pontosság adódott, ugyanakkor ezen paraméter növelésével trendszerű csökkenés volt megfigyelhető a kategorizálás helyességét a tesztadatokra vonatkozóan vizsgálva. (A legmagasabb arányban a $k = 4$ legközelebbi szomszédot megvizsgáló modell sorolta be helyesen a tesztadatbázis elemeit.)



5. ábra: A tesztadatbázisban szereplő díjmentesített szerződések helyes besorolási aránya az elkészített k-NN modellekben különböző k értékek mellett.

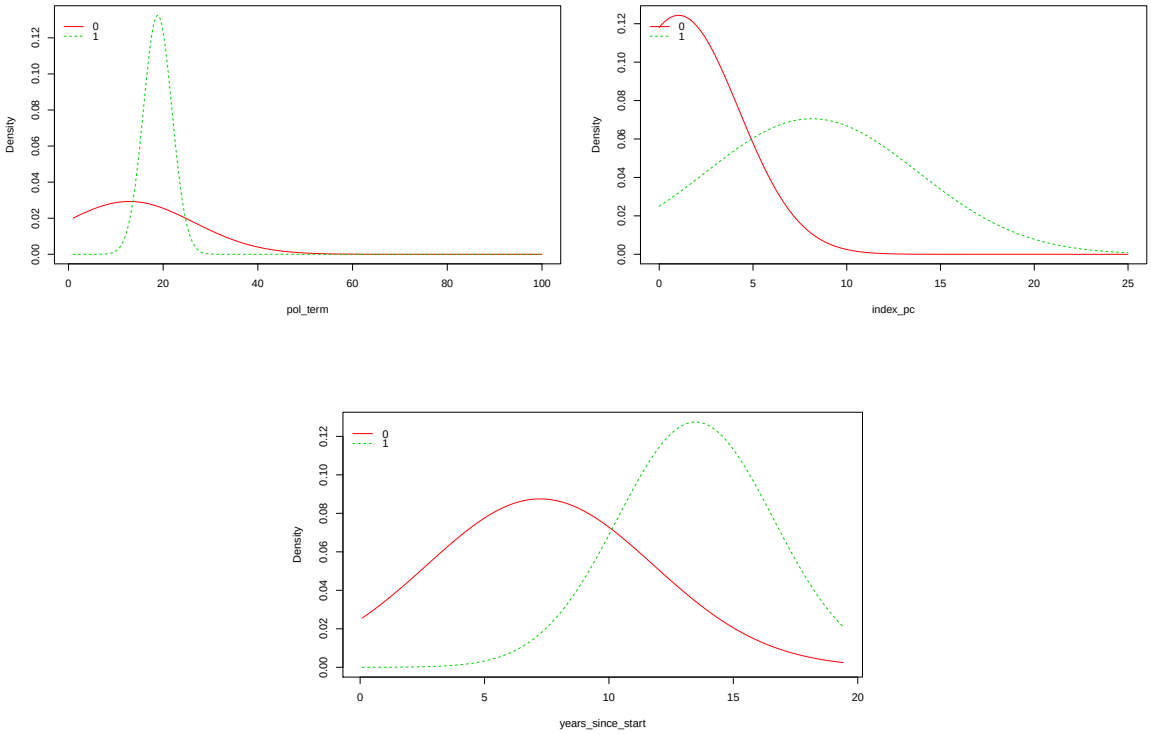
Mivel a végrehajtandó klasszifikációs feladat elsődleges célja a díjmentes leszállításon átesett szerződések azonosítása, érdemes megvizsgálni, hogy az előállított modellek célzottan erre a vonatkozásra fókuszálva hogyan teljesítettek (illetve a gyakorlati megközelítés alapján is kiemelten fontosnak nevezhető ezen szempont, hiszen egy biztosítót valószínűleg érzékenyebben érinti, ha egy olyan szerződésén történik díjmentesítés, amelytől ez nem volt várható előzetesen, mint ha potenciális díjmentesítőként tart számon egy olyan szerződést, amelyen végül nem következik be ilyen jellegű módosítás). Az 5. ábrán látható, hogy a k paraméter növekedésével a díjmentesített szerződések helyes besorolási aránya az összes szerződés kategorizálásához hasonlóan trendszerűen csökkent (kb. 92%-ról 78%-ra), viszont lényeges különbségnek nevezhető, hogy az adatbázis ezen szegmensében a $k = 1$ érték mellett adódott a legpontosabb kategorizálás.

Ezen modell szerint tehát annak eldöntéséhez, hogy egy új szerződésre várhatóan történik-e díjmentes leszállítási kérelem benyújtása, elegendő a megfelelő paraméterek szerint a hozzá leginkább hasonló modellpont megkeresése és ez alapján történő predikció adása. (Mindazonáltal meg kell jegyezni, hogy a k paraméter ilyen módon történő beállításával a modell rendkívül érzékeny lesz az egyedi új adatpontok beépítésére.)

4.2.4. Naiv Bayes-modell

Az R programcsomag *naivebayes* könyvtárának segítségével az előző alfejezetben létrehozott tanító adathalmazon végrehajtottam egy naiv Bayes-modell felépítését, szintén a már korábban bekövetkezett díjmentesítést jelölő *pup_dummy* változó osztályozási feladatának elvégzése céljából. A teljes adatbázis kb. 85%-át képező adatok alapján megbecsülésre kerültek a módszerben felhasználandó különböző feltételes valószínűségek: a numerikus változókra kapott tapasztalati sűrűségfüggvényekből is látható, hogy például a díjmentesített szerződéseket tipikusan egyenlőtlenebb tartam jellemzi, mint a többi állománybeli szerződést, illetve általában magasabb indexálási faktor került megállapításra velük kapcsolatban a szerződés kezdetén, továbbá jellemzően hosszabb ideje tartoznak a biztosító állományába, mint egy díjfizető szerződés (ld. 6. ábra).

A teszt adatbázis elemeire előállítva a naiv Bayes modellből származó predikciókat (az összes rendelkezésre álló prediktort felhasználva) megállapítható, hogy ezeken az



6. ábra: A tanító adatbázis elemeinek felhasználásával adódó tapasztalati sűrűségfüggvények a díjmentesített és a díjfizető állományra (zöld, illetve piros színnel jelölve) a szerződések tartamát, indexálási paraméterét és a szerződés kezdete óta eltelt időt figyelembe véve.

adatpontokon a ténylegesen díjmentesített szerződések azonosítását tekintve ezen módszer nagyságrendileg ugyanolyan jól teljesített, mint az elkészített k-NN modellek közül a legjobb: több mint 92%-os pontossággal került megállapításra a díjmentes leszállítás bekövetkezése. Ugyanakkor fontos különbségként jelentkezett, hogy az elsőfajú hiba gyakorisága, azaz a tévesen díjmentesítettként azonosított, de ténylegesen díjfizető szerződések száma nagymértékben felülmúlta az előző modellben tapasztaltak eredményeit, így a nem díjmentesített szerződéseket vizsgálva csupán kb. 74%-os pontosságú besorolást sikerült elérni (ld. 7. táblázat), így a végleges összesített helyes találati arány is 75% körül alakult.

A modell összesített pontossági besorolása javítható volt, amennyiben kizárólag a a logisztikus regresszióban szignfikánsnak bizonyult változók kerültek felhasználásra prediktorokként, hiszen így a modellpontok 93%-át sikerült megfelelő osztályba

	0	1
0	4000	1398
1	35	438

7. táblázat: Az output naiv Bayes-modell tévesztési mátrixa.

sorolni (ld. 8. táblázat). Ugyanakkor ez a pozitív irányú változás a díjmentesített szerződések felismerési arányában bekövetkezett jelentős mértékű csökkenéssel járt együtt: a teszt adatbázis ezen elemeit az így kapott modell 75%-os sikerességgel tudta helyesen kategorizálni. (Érdemes megjegyezni, hogy a 0-nak adódó relatív gyakoriságok helyettesítésére alkalmazható „threshold" paraméter módosításával a modellek outputjában jelentős mértékű változás nem volt elérhető.)

	0	1
0	5100	298
1	119	354

8. táblázat: A kizárólag a korábban szignifikánsnak minősített prediktorokat alkalmazó naiv Bayes-modell tévesztési mátrixa.

4.2.5. SVM

Az R programcsomag *e*1071 könyvtárát használva az input adatbázis numerikus változóinak segítségével három különböző modern SVM modellt építettem fel a díjmentesítés bekövetkezésére vonatkozó *pup_dummy* célváltozó által meghatározott bináris osztályozás elvégzése céljából.

Az első modellben input prediktorként feltüntettem az összes folytonos (intervallum szinten mért) változót radiális kernelfüggvény alkalmazása mellett ($\gamma = 0,2, C = 1$ beállítással), a kapott adatbázis kb. 85%-át tanító adathalmazként használva és a fennmaradó részen tesztelve a módszert a 9. táblázat által tartalmazott eredmények adódtak. Az adatokból látható, hogy ez a kezdeti modell több mint 95%-os összesített pontossággal tudta helyesen besorolni a teszt adathalmaz elemeit az egyes díjmentesítési kategóriákba, ugyanakkor ez a magas

találati arány nagyrészt a díjfizető szerződések túlnyomórészt helyes azonosításából származott (közel 99%-os pontosság), miközben a díjfizető szerződések csupán alig több mint 55%-os arányban kerültek megfelelően besorolásra.

	0	1
0	5344	54
1	212	261

9. táblázat: A kiinduló SVM modell tévesztési mátrixa.

A korábbi vizsgálatok alapján nem sziginfikánsnak bizonyult magyarázó változók eltávolításával kialakuló modern SVM modell az alábbi besorolási eredményeket alakította ki:

	0	1
0	5386	12
1	104	369

10. táblázat: A módosított SVM modell tévesztési mátrixa.

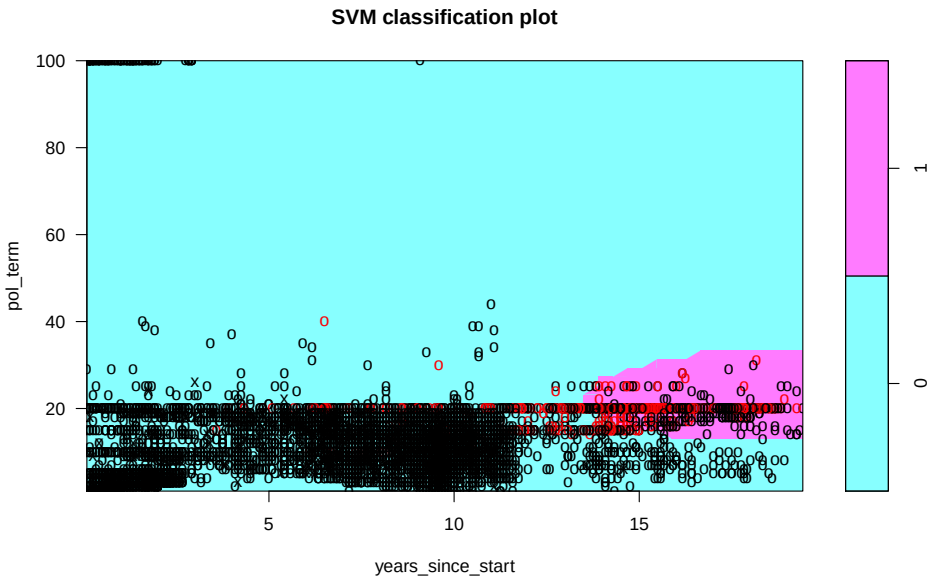
Megállapítható, hogy a kizárólag szignifikáns bemeneti változókat alkalmazó SVM modell jelentősen jobb teljesítményt nyújtott az előzőhöz képest: amellett, hogy a díjfizető szerződéseket szinte teljes pontossággal sikerült azonosítani, a díjmentes szerződések megfelelő kategorizálása tekintetében közel másfélszeres arányú javulás következett be (kb. 80%-ban pontos besorolás).

A harmadik SVM változatban az előzőekben ismertetett modellt módosítottam úgy, hogy radiális kernel helyett polinomiális magfüggvényt alkalmaztam ($\gamma = 0,2, c = 0, d = 3$ választással), az így kapott eredményeket a 11. táblázat tartalmazza. A táblázat adataiból látható, hogy bár a díjfizető szerződések azonosításának sikeressége kismértékben felülmúlta az előző, RBF-et használó modellbeli hasonló értéket, ugyanakkor a díjmentesen leszállított szerződéseket tekintve a polinomiális kernel jelentősebb arányban rosszabb teljesítményt nyújtott a második modellhez viszonyítva (kb. 75%-os pontosság).

	0	1
0	5393	5
1	124	349

11. táblázat: A polinomiális kernelt használó SVM modell tévesztési mátrixa.

A szignfikáns változókat és radiális bázisfüggvényt alkalmazó SVM 5 dimenziós prediktorterének egy kétdimenziós vetülete, illetve az ezen modell által megadott nemlineáris, „puha határt” alkalmazó szeparáció látható a 7. ábrán. Megállapítható, hogy ezen két változót vizsgálva az előállított modell a viszonylag régóta (legalább 14 éve) állományban lévő, egyben hosszú (kb. 15-35 év közötti) tartammal rendelkező szerződéseket helyezte a díjmentesített adatpontok kategóriájába.



7. ábra: A legjobb teljesítményt nyújtó SVM modell által megadott szeparáció a prediktortér egyik kétdimenziós vetületében.

5. fejezet

A szakdolgozat eredményeinek összefoglalása, következtetések

Szakdolgozatomban életbiztosítási szerződések díjmentes leszállítási szempontú karakterizációját végeztem el egy hazai biztosítótársaság által részemre elemzési céllal átadott adatbázis statisztikai és adatbányászati módszerekkel történő feldolgozása során. A díjmentesítés jelenségének általános szakirodalmi áttekintését követően a dolgozat fennmaradó része ezen szerződések megkülönböztető jellemzőinek azonosítására, a különböző statisztikai és adatbányászati modellek felépítésére, majd az adott adatbáziselemek klasszifikációjára, azaz „díjmentesített" és „díjfizető" kategóriákba való besorolására koncentrált. A dolgozat jelen fejezete az ezen eljárások segítségével kapott eredmények célzott módon való összefoglalását tartalmazza, majd általánosabb jellegű következtetéseket is megpróbálok megfogalmazni, természetesen arról sem megfeledkezve, hogy egyetlen biztosító egy évre vonatkozó adatainak alapján előállított eredmények nem feltétlenül általánosíthatók a szektor más társaságaira.

Az adatbázis 2011. év végéig díjmentesítésen átesett szerződésein belül a nemek közötti megoszlást, a tartam kezdetén választott fizetési gyakoriságot, illetve a tulajdonosok átlagos életkorát vizsgálva nem volt jelentős eltérés megállapítható a biztosítási állomány összességének paramétereihez viszonyítva. Ugyanakkor megfigyelhető volt, hogy a díjmentesített szerződések tipikusan az átlagosnál hosszabb tartammal rendelkeztek, valamint ezzel összefüggésben az átlagosnál szintén hosszabb ideje tartoztak a biztosító állományába. A díjmentesen leszállított szerződéseket jelentősen magasabb indexálási faktor jellemezte, mint az állomány

egészét, ugyanakkor (a díjmentesítés folyamatának definíciójából adódóan) érzékelhetően átlagosan alacsonyabb biztosítási összeg volt megfigyelhető ezen csoporton belül. Érdekességnek nevezhető, hogy a díjmentessé vált szerződések tulajdonosainak átlagéletkora a megkötés pillanatában kb. 4,5 évvel alacsonyabb volt, mint ugyanezen paraméter az összes szerződő körére kiterjedő módon vizsgálva (33,5 év, illetve 38 év).

A bináris logisztikus regresszió modelljének outputját elemezve megállapítható volt, hogy ezen eljárás szerint a díjmentesítési kérelem benyújtásának bekövetkezési valószínűségét az input változók közül kizárólag öt paraméter befolyásolta szignifikánsan, ezeken belül is legerősebben a szerződés megkötése óta eltelt évek száma és a szerződés tartama gyakorolta a legnagyobb hatást a díjmentesítés valószínűségére. A logisztikus regresszió klasszifikációs táblájának eredményeit vizsgálva pedig kijelenthető, hogy a felépített modell viszonylag elfogadható arányban volt képes besorolni az egyes szerződéseket a megfelelő kategóriákba, illetve a díjmentesített szerződések megfelelő kategorizálásában is kiemelkedően teljesített (86%).

A klasszifikációs probléma megoldására előállított döntési fák, illetve véletlen erdők szerkezetét vizsgálata során láthatóvá vált, hogy az optimális döntési fa a logisztikus regressziós modellben nem szereplő változókat (pl. nyereségtartalék összege) is felhasznált a besorolás elkészítése során, ugyanakkor csupán 72%-os pontossággal ismerte fel a díjmentesített szerződéseket. A különböző, outputként kapott véletlen erdők nagyságrendileg ugyanakkor a logisztikus regresszióhoz hasonló magas pontossági arányt tudtak elérni a klasszifikáció során (83-87%). A varImpPlot diagram alapján megállapítható volt, hogy az erdőket alkotó döntési fák felépítése során a biztosítási összeg, a szerződés tartama, az indexálás mértéke, illetve a szerződés kezdete óta eltelt idő nagysága bizonyultak a legfontosabb magyarázó változóknak.

A szakdolgozat elkészítése során megvizsgált három egyéb klasszifikációs módszer (k-NN, naiv Bayes modell, SVM) közös jellemzője, hogy egyfajta „fekete dobozként” működnek abból a szempontból, hogy az általuk kimenetként adott osztályozási modell felépítését kialakító tényezők az átlagosnál nehezebben megállapíthatók. Mindemellett az egyes k-NN modellek ($k = 1, \ldots, 100$) létrehozása során kizárólag a logisztikus regresszió által szignifikánsnak talált numerikus változók sztenderdizált értékeit alkalmaztam: a díjmentes szerződésekre koncentrálva a legmagasabb helyes

besorolási arányt a csupán az aktuális új modellponthoz ezen változók terében legközelebb eső szerződést figyelembe vevő modell adta outputként ($k = 1 : 92\%$).

Érdekes párhuzam volt megfigyelhető a naiv Bayes és az SVM modelleket vizsgálva: míg a díjmentesített szerződések helyes felismerési aránya az előbbi módszerben drasztikusan (92%-ról 75%-ra) csökkent áttérve a kizárólag a korábban szignifikánsnak bizonyult prediktorokat alkalmazó változatra, ezzel szemben a support vector machine modelleknél a nem szignifikáns változók elhagyása a díjmentesen leszállított szerződések sikeresen azonosított hányadának jelentős mértékű emelkedését eredményezte (55% helyett 80%).

A dolgozat eredményeit összegezve levonható a következtetés, hogy a vizsgált adatbázis elemeit tekintve számos módszer alapján a díjmentesített szerződéseket a biztosítási összeg, a tartam, a szerződéskötéstől számítva eltelt idő, illetve az indexálási faktor nagysága alapján lehetett leginkább megkülönböztetni a díjfizető állomány elemeitől. Megjegyzésre érdemes tény, hogy például a szerződők életkora, neme vagy az általuk választott fizetési gyakoriság tipikusan nem járult hozzá szignifikánsan a díjmentesen leszállított szerződések karakterizálására szolgáló predikciók érdemi javulásához. Ugyanakkor fontosnak tartom hangsúlyozni, hogy a legjobb klasszifikációs besorolási pontosságot elért módszerek (k-NN, véletlen erdők) helyes találati aránya sem nevezhető teljes mértékben kielégítőnek, hiszen még ezen modellek is kb. minden tizedik díjmentesített adatpont felismerését nem tudták kivitelezni. Így megállapítható, hogy egy nagyobb biztosítótársaság szempontjából is elfogadható, legalább 95%-os pontosságot eredményező modell létrehozásához a szakdolgozatban vizsgált változókon kívül nagy valószínűséggel szükséges lenne az egyes szerződésekre jellemző egyéb paraméterek (pl. a szerződő iskolai végzettsége, foglalkozása, háztartásának egy főre jutó havi jövedelme stb.) bevonása is.

Irodalomjegyzék

[1] Naomi S. Altman: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46. évf. (1992) 3. sz., 175–185. p.

[2] Banyár József: *Az életbiztosítás alapjai.* 1994, Bankárképző - Biztosítási Oktatási Intézet, 111–114. p.

[3] Banyár József: *Életbiztosítás.* 2003, Aula Kiadó, 230–233. p.

[4] Jan G. Bazan – Stanislawa Bazan-Socha – Sylvia Buregwa-Czuma – Lukasz Dydo – Wojciech Rzasa – Andrzej Skowron: A classifier based on a decision tree with verifying cuts. *Fundamenta Informaticae*, 143. évf. (2016) 1-2. sz., 1–18. p.

[5] Leo Breiman: Random forests. *Machine Learning*, 45. évf. (2003) 1. sz., 5–32. p.

[6] Leo Breiman – Jerome Friedman – Charles J. Stone – R. A. Olshen: *Classification and Regression Trees.* 1984, Wadsworth & Brooks/Cole Advanced Books & Software, 25–46. p.

[7] Corinna Cortes – Vladimir Vapnik: Support-vector networks. *Machine Learning*, 20. évf. (1995) 3. sz., 273–297. p.

[8] Nadine Gatzert: Implicit options in life insurance: An overview. In *Working Papers on Risk Management and Insurance No. 33.* (konferenciaanyag). 2009, 2–10. p.

[9] George H. John – Pat Langley: Estimating continuous distributions in Bayesian classifiers. *Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence*, 1995., 338–345. p.

[10] Kovács Erzsébet: *Többváltozós adatelemzés.* 2014, TypoTex Kiadó, 126–147. p.

[11] Mostafa Langarizadeh – Fateme Moghbeli: Applying naive Bayesian networks to disease prediction: A systematic review. *Acta Informatica Medica*, 24. évf. (2016) 5. sz., 364–369. p.

[12] Chih-Min Ma – Wei-Shui Yang – Bor-Wen Cheng: How the parameters of k-nearest neighbor algorithm impact on the best classification accuracy in case of Parkinson dataset. *Journal of Applied Sciences*, 14. évf. (2014) 2. sz., 171–176. p.

[13] Michael Reynaldo Phangtriastu – Jeklin Harefa – Dian Felita Tanoto: Comparison between neural network and support vector machine in optical character recognition. *Procedia Computer Science*, 116. évf. (2017), 351 – 357. p.

[14] Pröhle Tamás – Zempléni András: *Többdimenziós statisztika.* 2013, TypoTex Kiadó, 56–59. p.

[15] Tompa Krisztina Zsuzsa: A hagyományos életbiztosítási termékekben rejlő díjmentesítési opció értéke. *Biztosítás és Kockázat*, 3. évf. (2016) 1-2. sz., 62–85. p.