EÖTVÖS LORÁND TUDOMÁNYEGYETEM

TERMÉSZETTUDOMÁNYI KAR

BUDAPESTI CORVINUS EGYETEM

KÖZGAZDASÁGTUDOMÁNYI KAR

Virtás Dávid

Prediktív módszerek és gépi tanulás alkalmazásai a biztosításban

Biztosítási és pénzügyi matematika
MSc szakdolgozat

Témavezető:
Vékás Péter
Operációkutatás és Aktuáriustudományok Tanszék

Budapest, 2019

Tartalomjegyzék

Táblázatok jegyzéke

Ábrák jegyzéke

Köszönetnyilvánítás

Ezúton szeretném megköszönni konzulensemnek, Vékás Péternek, hogy megismertette velem az adatelemzésben rejlő szépségeket, valamint szakértelmével, magyarázataival és segítőkészségével lehetővé tette, hogy a dolgozat elkészüljön.

Köszönöm kollegámnak, Horváth Rolandnak, akihez bármilyen felmerülő probléma esetén fordulhattam tanácsért.

Természetesen köszönök mindent a családomnak, akik az egyetemi éveim alatt folyamatosan támogattak.

„If all you have is a hammer everything looks like a nail”

(Abraham Maslow)

1. Bevezetés

Egy életbiztosító számára az egyik legfontosabb kockázat fajta a törlési kockázat. Leginkább azért, mert egy szerződés felmondása azzal jár, hogy a biztosító elesik a további díjbevételtől, ami negatív hatással van az eredményére. A dolgozat elején a törlési kockázat mélyreható ismertetését végzem el. Bemutatom, hogy a szabályozó hatóság milyen intézkedéseket hozott annak érdekében, hogy a biztosítók szolvensek maradjanak és milyen szerepet játszik ebben a törlési kockázat. Ismertetem a törlések lehetséges fajtáit, melyet egy szakirodalmi összefoglaló követ. Ebben leírom a manapság legaktuálisabb részkockázat megjelenések történetét, és a jelenleg fennálló veszélyeket. Ez után biztosításmatematikai képletekkel, valamint egy számpéldán keresztül kiszámolom, hogy mennyit ér egy visszavásárlási opció. Mivel a közgazdasági és matematikai feltevések sokszor a valóságtól elrugaszkodottak, ezért fontosnak tartottam bemutatni egy más szemléletet is. Az elméleti összefoglaló legvégén viselkedésközgazdaságtani szempontból mutatom be a törlések lehetséges okait.

A dolgozat fő célja olyan prediktív modellek elkészítése, amelyekkel akár már az előzetes elbírálás során azonosítani tudják az életbiztosítók a kockázatosabb ügyfeleket. Az elemzésben olyan adatbányászati technikákat és gépi tanulási módszereket alkalmazok, amelyekkel reményeim szerint sikeresebben lehet a törléseket előre jelezni, mint a hagyományos módszerekkel. Ennek érdekében egy elméleti összefoglalóban ismertetem azokat a legfontosabb feltevéseket, összefüggéseket melyek elengedhetetlenek ilyen modellek elkészítéséhez.

Egy biztosítási szerződés életében – csak a törlési kockázatot figyelembe véve – három kockázatosabb időszak van, az első év, az az időpont, amikor az ügyfél már jogosulttá válik valamilyen visszavásárlási összegre, és az a pillanat, amikortól a visszavásárlást már nem bünteti a biztosító. A rendelkezésemre álló külföldi biztosító adatait arra fogom használni, hogy kiderítsem, mi befolyásolja a visszavásárlásokat ezekben az időpontokban.

1. A törlési kockázat

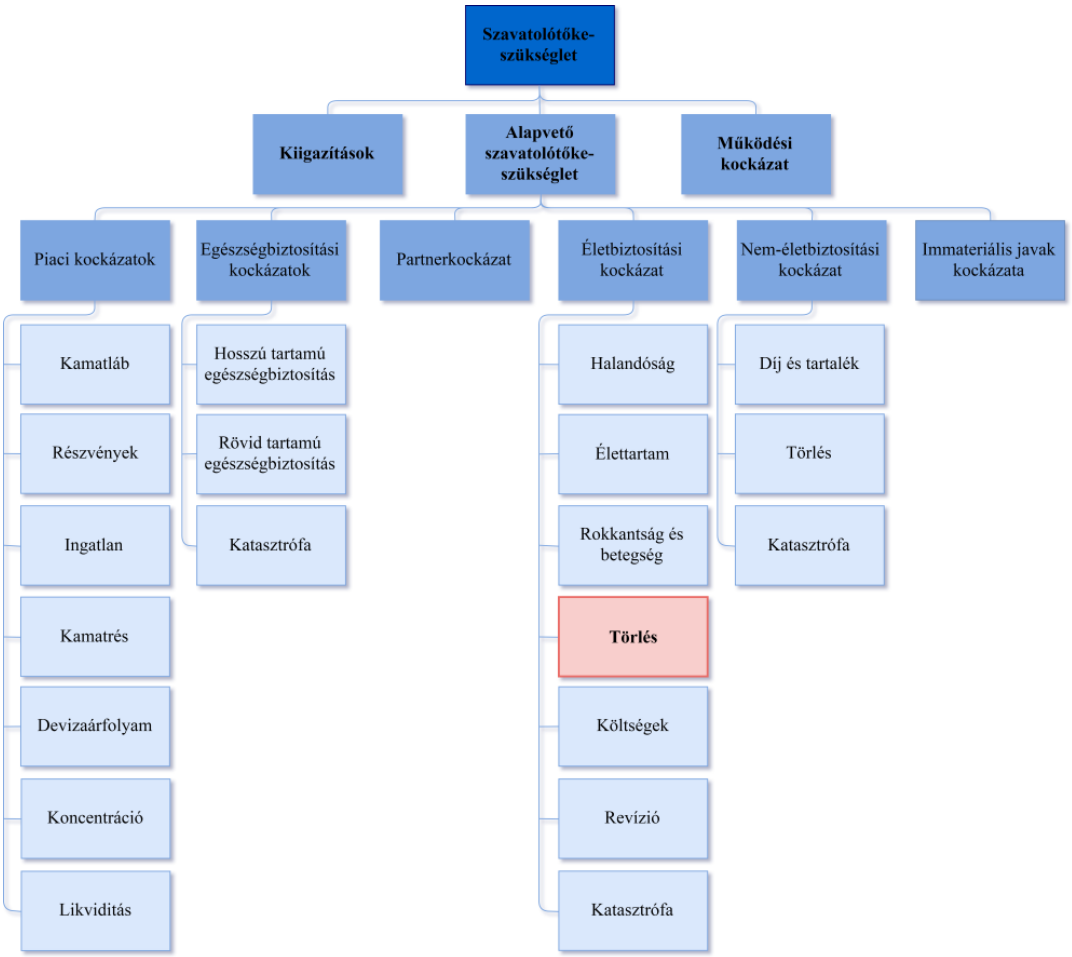
Ebben a fejezetben bemutatom a törlési kockázat Szolvencia II-ben meghatározott definícióját, kitérek arra, hogy milyen szerepet játszik a végleges szavatolótőke-szükséglet számolásában, valamint a fejezet második részében ismertetem a törlések fajtáit.

1.1 Törlés a Szolvencia II-ben

2016. január 1-től életbe lépett a Szolvencia II (EU, 2009), melynek első pillére szabályozza a kvantitatív követelményeket. Ennek az egyik része a szavatolótőke-szükséglet szabályozása, melyet úgy kell meghatározni, hogy egy éves távlatban, a biztosító minden fizetési kötelezettségének eleget tudjon tenni 99,5%-os valószínűséggel. Ez alapján minden biztosítónak – évente legalább egyszer – jelentést kell tennie a szabályozó hatóság felé. Kiszámítása három különböző módszerrel történhet:

- Standard Formula
- Belső Modell
- Részleges belső modell

A legnépszerűbb a standard formulával történő számolás melynek módszertanát a szabályozó hatóság határozta meg és attól eltérni – belső modell vagy részleges belső modell alkalmazásával – csak nagyon szigorú szabályoknak történő megfeleléssel lehet. Belső modell használata akkor lehet célszerű, ha a biztosító kockázati profilja nagymértékben különbözik a standard formula feltételezésitől. Ám a standard formulával történő számolást ekkor is be kell tudni mutatni a felügyelet felé.



1. ábra: A szavatoló tőke összetevői a Szolvencia II keretrendszerben (forrás: saját szerkesztés)

A szavatolótőke-szükséglet számolásához a biztosító állományát először hat modulra, majd ezeket további almodulokra kell osztani az 1. ábra szerinti kockázati megbontás alapján. A teljes szavatolótőke-szükséglet meghatározása az almodulok aggregálásával történik, figyelembe véve a közöttük lévő feltételezett korrelációk értékét és az almodulonként eltérő sokkokat. (Vékás, 2016) A dolgozatomban az életbiztosítási modul egyik almodulját vizsgálom, így az életbiztosítási modul szavatolótőke-szükségletét külön kiemelem:

$$SCR_{Life} = \sqrt{\sum_{i,j} Life_i * CorrLife_{i,j} * Life_j} \qquad (1)$$

$Life_i$, $Life_j$ egy-egy almodul, a közöttük lévő korrelációt $CorrLife_{i,j}$ jelöli. A korrelációs mátrix megtalálható a függelék 12. táblázatában. Hasonló módon történik az alapvető szavatolótőke-szükséglet (*BSCR*) számítása:

$$BSCR = \sqrt{\sum_{i,j} SCR_i * Corr_{i,j} * SCR_j} + SCR_{Imm} \quad (2)$$

Ahol az SCR_i és SCR_j egy-egy modul szavatolótőke-szükséglete, SCR_{Imm} pedig az immateriális javak kockázatából származó tőkeszükséglet. A modulok közötti korreláció szintén szükséges a számoláshoz, ezt $Corr_{i,j}$ jelöli és a korrelációs mátrix megtalálható a függelék 13. táblázatában.
A végleges szavatolótőke-szükséglet az alábbi képlettel számolható:

$$SCR = BSCR + Adj + SCR_{Op} \quad (3)$$

Ahol, *Adj* a tartalékok és az elhatárolt adók veszteségelnyelő hatásának korrekciója és SCR_{Op} a működési kockázat szavatolótőke-szükséglete.

A standard formulával számolt szavatolótőke-szükségletnek a legnagyobb részét a piaci kockázatok[1] teszik ki. Ezek közül a legnagyobbak a részvénypiaci, a kamatrés és a kamatláb-kockázati részmodulok, de ezen almodulok súlya nagyban függ a biztosítás típusától. A szavatolótőke-szükségletigény szerint a piaci kockázatokat az életbiztosítási kockázat követi, melynek a két legnagyobb a része a törlési- és a hosszú élet kockázat. (EIOPA, 2011) Mivel a dolgozatom témája a törlési kockázat, ezért a következő fejezetben ezt fogom részletesen bemutatni.

[1] Kamatláb-kockázat, részvénypiaci kockázat, ingatlanpiaci kockázat, kamatrés-kockázat, piaci koncentrációs kockázat, devizaárfolyam-kockázat

1.2 Törlési kockázat *(lapse)*

„A veszteség és a biztosítási kötelezettség értékében bekövetkező kedvezőtlen változás kockázata, amely a biztosítási szerződések törlési, megszüntetési, megújítási és visszavásárlási arányok szintjében vagy volatilitásában bekövetkező változásokból ered” (EU, 2009).

Ebben a fejezetben bemutatom a törlési kockázat összetevőit, valamint néhány szakirodalmat, amelyek a kockázat különböző aspektusait vizsgálva segíthetnek a megértésében.

Egy biztosítási szerződés törlése a következők valamelyikét jelenti: *díjmentes leszállítás*, *visszavásárlás*, *díjcsökkentés* vagy *szüneteltetés*, valamint a *harminc napon belül történő felmondás*.

1.2.1 Maradékjogok

A biztosító által beszedett díjtartalék egy későbbi szolgáltatás kifizetésének fedezetére szolgál, az ebből adódó elszámolási kötelezettség neve a maradékjog. Tehát, ha a szerződés valamilyen okból felbomlik, az ügyfél jogosulttá válik valamekkora kifizetésre, melynek alapja az említett díjtartalék. A maradékjogoknak két fajtája van az egyik a díjmentes leszállítás, a másik a visszavásárlás.

1.2.1.1 Díjmentes leszállítás

Ha az ügyfél már nem tudja/akarja fizetni a biztosítását, azonban nincs szüksége a díjtartalékra, akkor kérheti a biztosítót a szerződésének díjmentesítésére. Ebben az esteben a biztosító úgy tekint a meglévő díjtartalékra, mintha az ügyfél egy új, egyszeri díjas biztosítást vásárolt volna a díjtartalék összegéből. Az új biztosítás tartama megegyezik az eredeti biztosításból fennmaradó idővel. Lényegében az ügyfélnek nem kell több díjat fizetnie, cserébe a biztosítási összeg is lecsökken. (Banyár, 2016)

1.2.1.2 Visszavásárlás

A *visszavásárlás* egy életbiztosítási szerződésben foglalt opció. A biztosítási esemény bekövetkezése, vagy a tartam lejárta előtt, a szerződő megszűntetheti a szerződését. Ekkor jogosulttá válik egy előre meghatározott összeg kifizetésére, amelynek a mértéke a tartam alatt felhalmozódott díjtartaléktól függ. Ennek a bekövetkezése a biztosító cash-flowját

hosszútávon negatívan érinti, hiszen elesik a további díjbevételektől. Azonban rövidtávon nyereséget realizál a szerződésen, azért, mert az addig felhalmozott tartaléknak csak egy részét adja vissza a biztosítottnak. A tartalék és a kifizetett összeg különbségét nevezzük visszavásárlási büntetésnek. Ennek a célja, hogy a biztosítót kompenzálja az elmaradó haszonért, valamint arra ösztönzi az ügyfeleket, hogy ne töröljék le a szerződésüket, hiszen általánosan elmondható, hogy a biztosító arra törekszik, hogy minél több ügyfele legyen. (Banyár, 2016)

A visszavásárlások jelentik a legnagyobb törlési kockázatot. Nemcsak azért, mert ez a leggyakoribb módja a szerződés idő előtti megszüntetésének, hanem azért is, mert ez az egyetlen, ahol biztosítónak ténylegesen kifizetési kötelezettsége keletkezik. A dolgozatom második részében úgynevezett törlési (*lapse*) modelleket fogok készíteni, amikben a törlést, mint visszavásárlást értem.

1.2.2 Díjcsökkentés és szüneteltetés

A *díj csökkentése* és *szüneteltetése* két nem túl gyakori opció. Előbbi arra vonatkozó kérelem, hogy a biztosító csökkentse az ügyfél rendszeres díját, amivel arányosan természetesen a biztosítási összeg is csökken. Utóbbi egy meghatározott időre a biztosítási díj szüneteltetésére irányuló kérelem, szintén a biztosítási összeg rovására. (Banyár, 2016)

1.2.3 Törlés harminc napon belül

A *harminc napon belüli törlés* egy törvényben előírt joga az ügyfeleknek. A kötvényesítés[2] után harminc napig a biztosított felmondhatja a szerződését, ekkor a teljes befizetett díj visszajár számára, kivéve a kötvényesítés költségét. (Banyár, 2016)

[2]Az ajánlattétel után a biztosítónak tizenöt napja van a szerződést elfogadni vagy elutasítani. Amennyiben a szerződést elfogadja, megtörténik a kötvényesítés (a kötvény az ügyfélhez kerül), és ezt mind a két fél aláírásával igazolja.

2. A visszavásárlások szakirodalmi áttekintése

Az előző fejezetben bemutattam a törlések öt lehetséges fajtáit. Ebben a fejezetben részletesen ismertetem a visszavásárlások mögött rejlő kockázatokat. Bemutatom, hogy ez milyen kapcsolatban áll a hozamgörbével, és azt, hogy viselkedési közgazdaságtani szempontból hogyan racionalizálható egy biztosítási szerződés törlése.

2.1 Hozamgörbe kockázat

A garantált hozam[3], a többlethozam visszatérítés[4], a visszavásárlási opció csak néhány olyan opció és garancia, amelyek a biztosító kötelezettségei közé tartoznak. A fizetőképesség fenntartása érdekében ezen szerződéselemeket külön-külön érdemes értékelni. A biztosítók erre a ’90-es évek elején jöttek rá, azonban ekkor már túl késő volt a japán Nissan életbiztosítónak, ami 1997 októberében csődöt jelentett. A vesztét az okozta, hogy már nem volt képes kitermelni a 4,7%-os garantált hozamot ügyfelei számára, így hagyva hátra körülbelül 2,56 milliárd dollárnyi fedezetlen követelést. (Grosen & Lochte Jorgensen, 2000)

A fenti tanulmány azt vizsgálja, hogy miért kell, és hogyan lehet a fent említett három elemet külön-külön értékelni, valamint megemlíti, hogy a Nissan csődje előtt miért nem foglalkoztak ezzel. Az első indok az, hogy sokan egyszerűen nem voltak tisztában vele. A második, hogy nem foglalkoztak vele, nem érzékelték a hosszú távú szerződésekben rejlő kockázatot. A harmadik érv az, hogy akkor még nem állt rendelkezésre az az eszköztár, amivel az értékelés megtörténhetett volna. Ezt a problémát körül járva készült el a szerzőpár tanulmánya, ami rávilágít arra, hogy a biztosítási szerződések értéke erősen függ a piaci hozam és a garantált hozam különbségétől. Készítettek egy modellt, mellyel bizonyították, hogy a csőd valószínűsége nagymértékben csökkenthető, ha kevésbé

[3] Az a hozam, amit a biztosító már a szerződéskötéskor, de a tartam egészére, éves szinten garantál az ügyfeleinek.

[4] A technikai kamaton felül elért hozam egy részét a biztosító visszajuttatja az ügyfeleinek.

kockázatos eszközökbe fektetnek. Ezzel azonban az a probléma, hogy a kevésbé kockázatos termékeken kevesebb hozamot lehet elérni, ami visszautal az eredeti problémára.

(Fedoria & Förstemann, 2015) 2005 és 2013 között hatvan német biztosító adatait felhasználva készített elemzést, melyben megállapították, hogy a hozamgörbe 2,1 százalékpontos emelkedése drasztikusan növelné a törlések számát. Ennek az oka az elmúlt években tapasztalható alacsony hozamkörnyezet, ami érzékenyen érinti a biztosítókat, mert a korábban ígért magas(abb) kamatokat nehéz előteremteni. A probléma megoldásának egyik lehetséges módja hosszú lejáratú kötvények vásárlása, hiszen ezek normális piaci körülmények mellett magasabb kamatozással rendelkeznek, mint a rövidebbek. Ez egy jó stratégia lenne, ha a biztosítónak csupán akkor keletkezne kötelezettsége, ha a biztosított meghal, vagy lejár a szerződése. Azonban a legtöbb életbiztosítás rendelkezik visszavásárlási opcióval, ami miatt ez mégsem olyan egyszerű. A bonyodalmat az okozza, hogy a hosszú lejáratú kötvények növelik a portfólió módosított átlagidejét[5], ezáltal érzékenyebbé téve azt a hozamgörbe felfelé tolódására. Egy másik nehézséget jelenthet, ha a biztosító a törlések miatt hamarabb kényszerül eladni a kötvényeit, mint ahogy azt tervezte. Persze rendelkeznek likvid eszközökkel is, de magas törlés számnál nem biztos, hogy azok elegendő fedezetet nyújtanak, ekkor pedig kénytelenek eladni a befektetési céllal vásárolt eszközöket. Az előbb említett problémák már a ’90-es évek elejétől jelentkeztek, amire a megoldást a technikai kamatok csökkentése jelentette, ami azonban egy újabb kockázatot rejt magában. Ugyanis, ha a piaci környezet változik, és a kamatok emelkedni kezdenek, a biztosító, a hosszú lejáratú szerződések miatt nehezen tud reagálni. Ekkor előfordulhat, hogy az ügyfelek tömegesen elkezdik visszavásárolni a szerződéseiket, abban bízva, hogy máshol magasabb hozamot realizáljanak a befektetéseiken.

2.2 A visszavásárlás értéke

Fontos kiemelni, hogy a miért és milyen feltevések mellett szeretnénk meghatározni a visszavásárlás értékét. A következőkben bemutatásra kerülő fejezetben kívánom összekötni a 2.1 és a 2.3 fejezetet. Ugyanis, az előző részben azt mutattam be, hogy

[5] A módosított átlagidő azt mutatja, meg, hogy a hozamgörbe 1 százalékpontos emelkedése, milyen hatással van a kötvény árára.

racionálisan mikor érdemes a biztosítási szerződést, egy másik pénzügyi termékkel helyettesíteni. A következő (2.3) fejezetben pedig arról lesz szó, hogy milyen nem racionális döntések vezethetnek a visszavásárláshoz.

2.2.1 Az értékelési környezet

Ebben a fejezetben ismertetem azt az értékelési környezetet, amelyben a későbbiekben bemutatott képletet igazak. Természetesen a való életben egyik feltételezés sem állja meg a helyét, de a következőkben egyébként sem bármiféle matematikai alkalmazásról lesz szó. Azt fogom bemutatni, hogy (Bacinello, 2003) hogyan próbálja elméletileg különválasztani és valahogy számszerűsíteni a visszavásárlás értékét. Ezért arra van szükség, hogy a pénzügyi és a biztosítási piacok tökéletesen kompetitívek legyenek, ne létezzenek tranzakciós költségek és a piac arbitrázsmentes legyen, valamint a biztosítási ügynökök is racionálisan dolgozzanak. A halálozási kockázat a kifizetés idejét, míg a pénzügyi kockázat a kifizetés mértékét, valamint a visszavásárlási hajlandóságot befolyásolja. Ezen feltételek szükségesek a racionális visszavásárlás feltevéséhez. Így már a szerződés értéke egyértelműen összehasonlítható egy másik pénzügyi termékével.

2.2.2 A modell feltevései

A 2.2.1 fejezetben ismertetett értékelési környezett mellett két fontosabb feltevésre is szükség van, ahhoz, hogy a következő fejezetben bemutatásra kerülő képletek értelmezhetőek legyenek. Az első, hogy a halálozási valószínűség úgynevezett kockázatsemleges valószínűségi mérték alatt kerüljön megállapításra. Valós valószínűségnek nevezzük, ha a biztosító képes állományát olyannyira diverzifikálni, hogy ne legyenek kilengések a mortalitási rátában. Ekkor nincsen semmilyen mortalitási kockázat, tehát a halálozási valószínűség kockázatsemlegesnek tekinthető. Ezt a valóságban nehéz kivitelezni, ezért szokás a valós valószínűséget korrigálni egy biztonsági loading hozzáadásával.

Szükség van még definiálni a pénzügyi kockázatokat is. Ehhez (Cox, et al., 1979) ismert tanulmánya nyújt segítséget, melynek csak az ide tartozó részét ismertetem. A pénzügyi kockázat modellezése konstans hozamú, kockázatsemleges pénzügyi termékeken alapszik, melyeknek hozama sztochasztikusan fejlődik és létrehozható belőlük egy úgynevezett referenciaportfólió. Amely aztán egyértelműen egységekre *(unitok)* osztható

és ezeknek az egységeknek az ára egyértelműen meghatározza a referenciaportfólió hozamát. A korábban feltett arbitrázsmentesség miatt, biztosan létezik egy kockázatsemleges mérték, ami alatt minden pénzügyi termék ára martingál, ha a kockázatmentes hozammal kerülnek diszkontálásra.

2.2.3 A visszavásárlási opció árának meghatározása

(Bacinello, 2003) cikkében olyan életbiztosítások árazásával foglalkozik, amelyek tartalmaznak visszavásárlási opciót. Ezeket a termékeket egy amerikai típusú put opcióhoz[6] hasonlítja, mert visszavásárláskor, a szerződő eladja a szerződését a biztosítónak.

A tanulmány, a fenti feltevések mellett azt látja be, hogy egy ilyen termék ára három különálló részre tagolható*:* alapbiztosításra *(basic contract),* visszavásárlási opcióra *(surrender option),* valamint a többlethozam visszajuttatásához kapcsolódó opcióra. Ez utóbbit a cikk írója *participating option*[7]-nek nevezi, én innentől többlethozam opcióként fogom említeni. Alapbiztosítás alatt egy olyan szerződést ért, amelynél nem lehetséges sem a díjat csökkenteni, sem pedig a szerződést visszavásárolni a tartam alatt. Ezen kívül egy szokásos rendszeres díjas vegyes biztosítás, melynek ára az alábbi képlettel számolható:

$$U^B = C_1 A^{(r)}_{x:T|} = C_1 \left[\sum_{t=1}^{T-1} (1+r)^{-t}\, {}_{t-1|}q_x + (1+r)^{-T}\, {}_{T-1}p_x \right] \qquad (4)$$

A visszavásárlási opció nélküli termék árának kiszámítása az alábbi módon történik:

$$U^P = \sum_{t=1}^{T-1} \pi(C_t)\, {}_{t-1|}q_x + \pi(C_T)\, {}_{T-1}p_x = \frac{C_1}{1+\mu} A^{(\lambda)}_{x:T|} \qquad (5)$$

[6] Put opció: jog egy pénzügyi termék eladására. Az amerikai típus azt jelenti, hogy az opció tulajdonosa a szerződés tartama alatt bármikor élhet ezzel a jogával.

[7] Participating option: A név a *participating policy*-ból ered, mely egy olyan termék, ahol a szerződő részesedik a többlethozam valamekkora részében

A többlethozam opció ára megkapható (5) és (4) különbségeként, azaz egy többlethozam opcióval rendelkező, de visszavásárlási opcióval nem rendelkező termék ára csökkentve az alapbiztosítás árával.

$$B = U^P - U^B = \\ = C_1 \left\{ \sum_{t=2}^{T-1} (1+r)^{-t} [(1+\mu)^{t-1} - 1]_{t-1|} q_x \right\} + \\ + C_1 \{ (1+r)^{-T} [(1+\mu)^{T-1} - 1]_{T-1|} p_x \} \tag{6}$$

A biztosítási összeg sztochasztikus fejlődését is feltételezve ($\{C_t,\ t = 1,\ 2, \dots,\ T\}$), a szerződés teljes értéke felírható $(n+\ 1)$-nominális fa átlagaként, ahol a fa gyökere a C_1, és mindegyik részfának n+1 ága van. Mivel a fa összeölelkező ezért az $(n+1)^t$ lehetséges eset helyett $\binom{n+t}{n}$ különböző ága van. A fa leveleitől a gyökérig visszafelé haladva meghatározható a szerződés teljes értéke. Ebből kivonva (5)-öt, megkapható a visszavásárlás értéke. A fejezetben használt változók jelentése megtalálható a függelék 14. táblázatában.

2.2.4 Számpélda

A tanulmány egy számpéldán keresztül bemutatja a fent leírt módszert, az arányok ismertetése miatt röviden ismertetem a szerző eredményeit.

Az alábbi paraméterek rögzítettek: A *halálozási adatok* az 1991-es olaszországi női halálozási táblából, *öt éves szerződés*, évenként *kétszázötven kereskedési nappal* (ezen napokon változik a referenciaportfólió értéke) és tízezer dolláros kezdő *biztosítási összeg*. A visszavásárlási opció értékét a többi paraméter változtatásával számolta ki, ezen eredmények megtalálhatóak az eredti cikk 2-8. táblázatában. Egy konkrét példát kiemelnék: $x = 50, r = 0.05, i = 0.02, \eta = 0.5\ \sigma = 0.15, \rho = 0.035$ mellett a visszavásárlási opció értéke 128 dollár (Bacinello, 2003).

2.3 Viselkedés közgazdaságtani szemlélet

(Campbell, et al., 2014) a szakirodalomban sokszor előforduló két olyan hipotézist nevez meg amelyek magyarázzák a törlések jelentőségét. Az egyiket *„hozamgörbe*

hipotézisnek”, a másikat *„vészhelyzeti alap hipotézisnek”* nevezi. Előbbi a hozamgörbe emelkedésével egyidejűleg a törlések emelkedését feltételezi, míg utóbbi alatt a szerződő fél pénzügyi nehézsége esetén jelentkező törlést, mint tisztán pénzügyi döntést említi. Ezek mellett négy viselkedés közgazdaságtani szempontot ismertet. Ezek a döntés-egyszerűsítés *(decision shortcuts),* az értékbecslés, valamint az érzelmi és társadalmi hatások.

2.3.1 Döntés-egyszerűsítés

A 2.2.1 és a 2.2.2 fejezetben azokat a feltevéseket ismertettem, ami ahhoz szükséges, hogy matematikai modellekkel jobb képet lehessen kapni a törlésekről. Azonban (Campbell, et al., 2014) tanulmánya éppen arról szól, hogy kevés olyan szerződő van, aki ehhez hasonló módon hozza meg pénzügyi döntéseit. Sokkal inkább egyéb, nem racionális döntéseket hoznak, amikor úgy döntenek, hogy a szerződésüket törlik. Ezeket nevezik *döntés-egyszerűsítéseknek*, és a következő alfejezetekben ezeknek a különböző fajtáit mutatom be.

2.3.1.1 Relatív választások

Két pénzügyi termék összehasonlításának a racionális módja, a termékek külön-külön értékelése, majd ezen értékek összehasonlítása. Azonban a legtöbben inkább az alapján választanak, hogy a két termék közül melyik tűnik jobb üzletnek. E döntésében sokszor szerepet kapnak egyéb, nem feltétlen racionális és valamelyik irányba elfogult pszichológiai tényezők. Ezért fontos a biztosítónak a róla kialakított kép, vagy például egy termék eladásánál megcélzott célcsoport.

2.3.1.2 Mentális könyvvitel

Az emberek egy része fejben kialakít egy költségvetést, amiben a kiadásaikat kategorizálják. Ezt nevezzük mentális könyvvitelnek *(mental accounting)*. A törlések vizsgálatánál elkülöníthető két különböző típusú mentális könyvvitel. Az egyik, hogy miként tekintenek az ügyfelek a saját biztosításukra. Vannak, akik a szerződés díját kiadásnak, és vannak, akik megtakarításnak tekintik. A másik, hogy a szerződő mire szeretné felhasználni a szerződésen gyűjtött összeget: jövőbeni kiadását tervezi fedezni belőle, vagy csak tőkefelhalmozás céljából vásárolt életbiztosítást. A mentális könyvvitel magyarázza azt is, hogy két technikailag azonos biztosításon más törlési arány mutatkozik, akkor, ha biztosításként vagy ha megtakarításként kerül a piacra.

2.3.1.3 Értékbecslés

Viselkedési közgazdaságtani tanulmányok megmutatták, hogy az érték egy relatív fogalom, amit az érzelmek könnyen befolyásolhatnak. Hiperbolikus diszkontálás alatt értjük azt a jelenséget, amikor az egyének úgy cselekednek, mintha az időben távolabbi pénzáramlást egy magasabb hozamgörbével diszkontálnák, mint a közelieket. Tehát az ügyfelek a díjat egy anyagi tehernek fogják fel, és nem gondolják, hogy a szerződésük hosszú távon kifizetődő. Ezt már a biztosítók is felismerték és ennek köszönhető a magasabb díj, egy gyakoribb díjfizetésű szerződésen. Hiszen ezek az ügyfelek többször kerülnek döntéskényszerbe a díjfizetés kapcsán. Természetesen a gyakoribb díjfizetéshez magasabb költségek tartoznak, ami egy másik ok amiért ezen szerződések díjai drágábbak. Mindezt alátámasztja (Eling & Kiesenbauer, 2013) akik megmutatták, hogy a magasabb díjfizetési gyakorisághoz magasabb törlési arány párosul.

2.3.1.4 Érzelmi hatások

Az egyik fő ok, amiért az emberek életbiztosítást vásárolnak az, hogy kockázatkerülők. Ha az ügyfelek biztosítási érdeke változik (például felnőtt lesz a gyermekük), akkor változhat a hozzáállásuk is ezzel kapcsolatban. Másik ok az, hogy egyfajta pénzügyi kontrollra van szükségük. Erre egy rendszeres (általában havi) díjas életbiztosítás egy tökéletes eszköz, hiszen ez egy kötelezettség a szerződőnek, ami arra „kötelezi", hogy a jövedelme valamekorra részét megtakarításra fordítsa.

2.3.1.5 Társadalmi hatások

Az egyének döntéseit társadalmi hatások is befolyásolják, például a család vagy a barátok véleménye. Így csak ritkán fordul elő, hogy a biztosításokkal kapcsolatos döntést egyedül hozzák meg az emberek. Ezt nevezik *utánfutó-hatásnak.*

Látható tehát, hogy a biztosítási szerződések törléseinek modellezésével már hosszú ideje rengetegen foglalkoznak. A törléseknek többféle hatása is van a biztosító eredményére. Előfordulhat, hogy olyan hamar következik be, hogy már a kezdeti költségeket (szerzési, kockázat-elbírálási költségek) sem térülnek meg. Sőt bizonyos kontraszelekciós hatása is van, hiszen megfigyelhető, hogy azok az ügyfelek, akiknek a tartam folyamán jobban növekszik a halálozási vagy rokkantsági kockázata kevésbé hajlamosak a törlésre. Ezzel szemben azok, akiknek ezek a kockázati tényezők kevésbé növekednek, inkább hajlamossá válnak egy esetleges visszavásárlásra. Ez ahhoz vezet, hogy portfólió szinten a biztosítónak magasabb fizetési kötelezettsége keletkezik, mint amivel tervezett. A

harmadik fontos szempont a likviditási kockázat. Magas törlésszám esetén előfordulhat, hogy a biztosító nem tud eleget tenni a vállalt kötelezettségeinek, ezért el kell adnia az eszközeit. Egy eszköz tervezettnél korábbi értékesítése pedig akár nagyon magas vesztességgel is járhat. Ha tehát a biztosító nem tudja kellően sikeresen megbecsülni, hogy mikor, milyen eszközöket kell eladnia, akkor előfordulhat, hogy a nagyobb veszteség elkerülése miatt a kellőnél likvidebb befektetéseket választ, ami viszont a hozam kárára megy. Extrém helyzetekben a biztosító hírneve is romolhat, ami így még több törléshez vezethet. Azonban az alacsony törlésszám is veszélyes lehet a biztosító eredményére, akkor, ha nem tudja kitermelni a garantált hozamot. (Kuo, et al., 2003)

3. Az alkalmazott módszerek elméleti bemutatása

3.1 Logisztikus regresszió

A logisztikus regresszió a biztosítás és pénzügyi matematika mesterszak törzsanyagának részét képezi, ezért csak egy rövid emlékeztetőként említem meg, hiszen ez az egyik legismertebb klasszifikációs eljárás. Nagyon hasonló a lineáris regresszióhoz, azonban a lineáris kapcsolat nem az eredményváltozó és a prediktorok között áll fent, hanem az eredményváltozó és az úgynevezett log odds között.

$$logit(y) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k \beta_k \qquad (7)$$

ahol *p* jelöli egy adott kimenet valószínűségét. A becslés *p*-re a következőképp kapható meg:

$$\hat{p} = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k \beta_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k \beta_k}} = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\cdots+\beta_k \beta_k)}} \qquad (8)$$

3.2 Döntési fák, véletlen erdők

3.2.1 Döntési fa

A döntési fa egy széles körben elterjedt algoritmus, amellyel egyszerű lépések segítségével hozhatunk meg bonyolult döntéseket. Legnagyobb előnye, hogy könnyen vizualizálható és értelmezhető, valamint klasszifikációs és regressziós problémákra egyaránt alkalmazható. A legnagyobb hátránya az, hogy hajlamos a túlillesztésre (Bodon, 2010). Habár a két módszertan nagyon hasonló, csupán abban térnek el, hogy az előbbinél egy kvalitatív, míg utóbbinál egy kvantitatív eredményváltozó becslése a cél. Ebben a fejezetben a klasszifikáció módszertanát mutatom be.

Klasszifikációs fa alkalmazásakor az első lépésben a megfigyeléseket $(X_1, X_2, \dots, X_p)$, J darab diszjunkt halmazba osztjuk: R_1, R_2, …, R_J. A halmazoknak elméletileg bármilyen alakja lehet, a gyakorlatban azonban többdimenziós téglalapokat szokás használni, így a végső modell könnyebben értelmezhető. A csomópontokon csak egy változóról és csak kisebb/nagyobb tulajdonság vizsgálatára kerül sor, így a végeredmény könnyen értelmezhető. Egy elem becslése (törlik-e a szerződést vagy nem) a halmazon belül leggyakrabban előforduló elem lesz. A fa „növesztése" egy rekurzív algoritmussal történik, úgy, hogy a becslés hibája a lehető legkisebb legyen. A hiba nagyságát a regressziós esetben az SSE (*átlagtól vett eltérések négyzetösszege*) mutatja meg. Ez klasszifikációs fánál értelemszerűen nem használható, logikus helyettesítése lenne az úgynevezett klasszifikációs hiba arány[8] (*classification error rate*). Azonban ez a mutató nem érzékeny a fa nagyságára, ezért csak a végső modell helyességét vizsgáljuk vele. A fa metszéséhez a gyakorlatban leggyakrabban a Gini indexet szokás alkalmazni.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{9}$$

[8] Klasszifikációs hiba arány: Azon elemek száma, amelyek nem tartoznak abba a halmazba, amelyikbe a modell sorolta és az összes elem hányadosa.

A Gini index a *K* kategória közötti teljes varianciát mutatja meg. A p_{mk} az *m*-edik csoport *k*-adik osztályának részaránya. Kis értéke azt jelzi, hogy egy csoporton belül magas a homogenitás. (James, et al., 2017)

A döntési fák egyik sajátossága, hogy hajlamosak a túlillesztésre. Ez azt jelenti, hogy létrejövő fa terebélyes lesz, azaz sok ága képződik. Ekkor a levelek szinte teljesen homogének, akár az is előfordulhat, hogy a modell tökéletesen illeszkedik a mintára, de egy másik sokaságon már sokkal rosszabb eredményt adna. A probléma a fák metszésével (*pruning*) kezelhető. Erre két bevált módszer létezik, az előnyesés (*pre-pruning*) és az utónyesés (*post-pruning*). Előnyeséskor a fa nem nő tovább, ha az újonnan létrejövő ág nem érne el, egy előre megadott szignifikanciaszintet. Általában ez egy χ^2 teszt, ahol a nullhipotézis az, hogy az új vágással létrejövő csoport független egy már meglévőtől. Utónyeséskor, az adatokat három részre kell osztani: *tanuló*, *keresztvalidáló* és *tesztelő* sokaságra. Először a tanuló adatokon létrejön egy terebélyes fa. Ennek a fának minden egyes részfájából egyenként levél képződik, majd az így létrejött fa, az eredeti fával összevetésre kerül. Ha az új fának kisebb a klasszifikációs hiba aránya, akkor a vágás maradandó. (Patel & Upadhyay, 2012)

3.2.1.1 Bagging

A döntési fák egyik problémája, hogy sok esetben rosszabbul teljesítenek, mint egy hagyományos regresszió, azonban létezik egy módszer, amivel növelhető a predikció pontossága. Ezt a módszert nevezzük véletlen erdőnek. A véletlen erdő eljárás a *bagging* eljárásból fejlődött ki, ezért először azt fogom bemutatni, majd utána térek rá a két módszer közti különbségre.

A *bagging* alapvető célja egy statisztikai tanuló eljárás varianciájának csökkentése. A neve a *bootstrap aggregation* kifejezés összeolvasásából ered. Az ötletet az az alapvető összefüggés ihlette, hogy *n* db független, azonos eloszlású valószínűségi változó átlagolásával, az átlag varianciája csökken. A variancia ilyen formájú csökkentéséhez arra lenne szükség, hogy a populációnak sok különböző mintáján alkossunk modelleket, majd a végső predikciót egy többségi szavazással döntjük el (regresszió esetén átlagolással). Általában egy adott minta áll a rendelkezésünkre, ezért ez a lehetőség a legtöbbször nem adott, helyette célszerű *bootstrappelni*[9]. A *bagging*-et *B* darab bootstrappelt mintán hajtjuk végre, úgy, hogy a fákat hagyjuk szabadon nőni és nem is metsszük azokat. Ekkor

[9] Bootstrap: visszatevéses mintavétellel, több különböző minta létrehozása.

B darab predikció a rendelkezésünkre áll, amiből többségi szavazással határozhatjuk meg, a végleges előrejelzést. Ezzel az eljárással pontosabb, azonban sokkal nehezebben értelmezhető eredményt kapunk, mint döntési fákkal. Az eredmények értelmezésében segíthet, ha megvizsgáljuk azt, hogy egy-egy változó átlagosan mennyire csökkentette a változó Gini indexét (*mean decrase in Gini index*). (James, et al., 2017)

3.2.2 Véletlen erdő

A véletlen erdő eljárás a *bagging* egy továbbfejlesztettje, amelynek célja a változók közötti korreláció csökkentése. Erre akkor van szükség, ha a változók közül egynek kiemelkedően magas magyarázó ereje van. A *bagging* eljárás alkalmazásakor, az előbb említett változó a legtöbb fának része lesz, így a kialakult fák hasonlóak lesznek, ami nem kívánt korrelációhoz vezet. Véletlen erdő használatával ez a korreláció csökkenthető, méghozzá úgy, hogy a döntési fák felépítése közben, a vágási pontokon a szokásos *p* darab változó helyett, csak véletlenszerűen választott *m* darab változó alapján engedjük meg a vágást. Általában *m* értéke közelítőleg megegyezik *p* négyzetgyökével. (James, et al., 2017)

3.3 Támaszvektor-gépek (SVM)

Ebben a fejezetben a support vector machine (SVM) algoritmust mutatom be, amely klasszifikációra, regresszióra, de még akár klaszterezésre is alkalmazható. Lineáris SVM-et már a ’60-as években használtak, népszerűsége azonban a modern SVM kialakulása után kezdett növekedni. Azóta ezzel a módszerrel jelentős eredményeket értek el. Többek között használható arc- és kézírásfelismerésre, képek klasszifikációjára, valamint segítséget nyújt a rákos sejtek feltérképezésében. (Dataflair Team, 2017) Fontos megkötés, hogy csak numerikus prediktorokat tud kezelni, valamint a kimeneti változó értéke csak -1 vagy 1 lehet, de előnye, hogy nem használ semmilyen véletlent. Az ebben a fejezetben bemutatásra kerülő SVM-ek alapkérdése megegyezik: hogyan lehet a prediktorok terét egy hipersíkkal elválasztani.

3.3.1 Lineáris SVM

A kiinduló feladatban a változók $\mathbb{R}^p$ terében egy p-1 dimenziós hipersíkot keres, mely tökéletesen elválasztja a kimeneti két kategóriáját egymástól.

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{10}$$

Ha az elválasztás lehetséges, akkor a hipersík felírható (10) formájában, és egy x^* megfigyelésről egy egyszerű behelyettesítéssel eldönthető, hogy a hipersík melyik oldalán helyezkedik el. Az így kapott szám nem csak azt dönti el, hogy x^* melyik csoportba kerül klasszifikálásra, hanem azt is, hogy ez a klasszifikáció mennyire megbízható. Minél nagyobb, annál biztosabb. Probléma azonban, hogy ha a térnek létezik egy ilyen felosztása, akkor végtelen létezik. A megoldáshoz a súlyvektort (**w**) a normalizálni kell, (**w**/||**w**||) ekkor a hipersíkhoz legközelebb eső pontokra (*támaszvektorok*) igaznak kell lenni vagy (11)-nek, vagy (12)-nek.

$$\mathbf{w} \cdot \mathbf{x} + b = +1 \tag{11}$$

$$\mathbf{w} \cdot \mathbf{x} + b = -1 \tag{12}$$

A feladat a hipersík és támaszvektorok közötti távolság (γ) maximalizálása. (γ)-ról belátható, hogy ebben az esetben az értéke $\frac{1}{\|\mathbf{w}\|}$,így tehát (γ) maximalizálása helyett $\|\mathbf{w}\|$ minimalizálása a feladat.

$$\min_{\boldsymbol{w},b} \|w\| : \\ y_i(w \cdot x_i - b) \geq 1 \qquad (i = 1,2,\dots,n) \tag{13}$$

3.3.2 Nemlineáris SVM

A gyakorlatban a két csoport a legtöbbször nem szeparálható lineárisan, amire megoldást a nemlineáris SVM jelenthet. Ekkor a prediktorokat nemlineáris transzformációval magasabb dimenziójú térbe transzformálva, már nagy valószínűséggel lineárisan szeparálható teret kapunk. Ez az úgynevezett Cover-tétel. (Cover, 1965)

Ekkor az előbbi optimalizációs probléma a következőre módosul, ahol ϕ a fent említett transzformációs függvény.

$$\min_{\boldsymbol{w},b} \|w\| : \\ y_i(w \cdot \phi(x_i) - b) \geq 1 \qquad (i = 1,2,\dots,n) \tag{14}$$

Ezzel azonban az a probléma, hogy az új dimenzió száma nincs korlátozva, ezért könnyen túl magas dimenzióba kerülhet a transzformálás, ami nagyon megnöveli a számítási igényt. Ezért célszerű bevezetni egy magfüggvényt (*kernelt*). A magfüggvény az eredeti tér két vektorára megadja azok új, transzformált térben vett kapcsolatát.

Név	$\kappa(\mathbf{u},\mathbf{v})$
Lineáris	$\mathbf{u} \cdot \mathbf{v}$
Polinomiális	$\gamma(\mathbf{u} \cdot \mathbf{v} + c_0)^d$
Radiális	$\exp(-\gamma\|\mathbf{u} - \mathbf{v}\|^2)$
Szigmoid	$\tanh(\gamma\mathbf{u} \cdot \mathbf{v} + c_0)$

1. táblázat: A nemlineáris SVM-nél alkalmazott gyakori magfüggvények

Kernelek használata azért lehetséges, mert a hipersík egyenletében csak azokat a pontokat kell figyelembe venni, amik a támaszvektorokat határozzák meg.

3.3.3 Modern SVM

A két előző SVM módszer tökéletes szeparációra törekszik, azonban ez nem mindig lehetséges, ha pedig lehetséges, akkor sem biztos, hogy a kívánt eredményhez vezet. Előfordulhat, hogy egyes megfigyelések túlsúlyosak lesznek, ezáltal a hipersík kialakításakor a határok torzulnak, ez pedig félrevezetheti a klasszifikációt. Ennek a problémának a leküzdésére találták ki a modern SVM-et, ami nem törekszik tökéletes elválasztásra, így néhány megfigyelés a hipersík rossz oldalára is kerülhet. Ezt nevezzük puha határnak *(soft margin)*. A modern SVM a puha határral kiegészített nemlineáris SVM. A határsértések nagyságát a *C* paraméterrel lehet beállítani, értékét célszerű keresztvalidációval meghatározni.

A probléma ebben az esetben az alábbi módon írható fel:

$$\begin{aligned} &\min_{\boldsymbol{w},b,\xi} \|w\| : \\ &y_i(w \cdot \phi(x_i) - b) \geq 1 - \xi_i \qquad (i = 1,2,\dots,n) \\ &\xi_i \geq 0 \qquad (i = 1,2,\dots,n) \\ &\sum_{i=1}^{n} \xi_i \leq C \end{aligned} \tag{15}$$

Ahol az ξ_i-k mutatják a határsértés mértékét. Ha $\xi_i = 0$, akkor az i-edik megfigyelés esetében nem történik határsértés, ha $0 < \xi_i < 1$, akkor a határmesgyében van, ha pedig $\xi_i > 1$, akkor a határ rossz oldalán van. (Shalev-Shwartz & Ben-David, 2014)

3.3.4 k-legközelebbi szomszéd (k-NN)

A k-legközelebbi szomszéd (*Nearest Neighbour*) eljárás egy nem-paraméteres tanulási eljárás. A nevét onnan kapta, hogy egy új minta klasszifikációjakor az ahhoz legközelebb álló *k* db megfigyelés egyszerű többségi „szavazása” alapján dönti el, hogy melyik csoportba klasszifikálja. Két fontos, technikai beállítást kell meghatározni: milyen távolság metrikát használjunk és mennyi legyen *k* értéke. A leggyakrabban használt távolsági mérték az euklideszi és a negatív koszinusz hasonlóság, de ezek mellett meg kell említeni még a Csebisev, Mahalabonis, és a Hamming távolságot is. Ezen távolságok kiszámítása módja megtalálható a függelékben, a 15. táblázatban. Azt, hogy melyiket célszerű használni el lehet dönteni találgatással, de létezik egy módszer, amely alapján az adatok segítségével ez a mérték tanulható.

k értékének kiválasztására nincs pontos matematika mérték, ami mindig a legjobb választást adja. Hüvelykujj szabály alapján gyakori választás a minta elemszámának négyzetgyöke. Egy másik lehetőség a könyökpont módszer[10] alkalmazása. Ezen kívül érdemes kiemelni az úgynevezett legközelebbi szomszéd módszert (k = 1 esetén). Ennek a módszernek az alkalmazásakor az alapján klasszifikáljuk az új elemet, hogy a tanuló adatbázisban mi áll hozzá a legközelebb. Ez egy logikus választásnak tűnik, de a gyakorlatban nem elég robusztus és gyakran felmerül a túlillesztés problémája. (Burkov, 2019)

3.3.5 Naív Bayes

A Naív Bayes a maximum likelihood becslést és a Bayes-tételt együttesen használja a változók klasszifikálására. Ekkor az új megfigyelést abba a csoportba érdemes sorolni, ahol a $P(Y_i|X)$ maximális. Y_i jelöli a megfigyelés i-edik osztályba tartozását $(Y = y_i)$, míg X a prediktorok értékeinek előfordulását $\left(X = \left(x_1, x_2, \dots, x_p\right)\right)$ a Bayes-tétel alapján:

$$P(Y_i|X) = \frac{P(X, Y_i)}{P(X)} = \frac{P(X, Y_i)P(Y_i)}{P(X)} \tag{16}$$

Mivel $P(X)$ mindig konstans, ezért a maximalizálásban nem vesz részt, $P(Y_i)$ ha nem adott, akkor a mintából a relatív gyakorisággal becsülhető. $P(X|Y_i)$ becsléséhez szükség lesz arra a feltételezésre, hogy egy osztályon belül az attribútumok feltételesen függetlenek egymástól. Ez azért fontos, mert a becsülendő paraméterek száma így jelentősen lecsökken. A feltételezés előtt ez a szám $l(2^k\text{-}1)$, utána csak $l \cdot k$ (l jelöli Y lehetséges értékeit, k bináris magyarázó változók számát).

Ha tehát a fenti feltételezés teljesül, akkor $P(X|Y_i)$ felírható az alábbi módon:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k|Y_i) = \prod_{j=1}^{k} P\left(X_j = x_j\middle|Y_i\right) \tag{17}$$

[10] Könyökpont módszer: Minden k esetén kiszámoljuk a külső szórásnégyzet és a teljes szórásnégyzet hányadosát, majd ezeket az értékeket ábrázoljuk. Ahol a görbében törés mutatkozik, azt az értéket választjuk k-nak.

Amennyiben X_j kategorikus változó akkor, a $P(X_j = x_j|Y_i)$ a mintából becsülhető a relatív gyakorisággal. Ha X_j folytonos változó, akkor a becsléshez ismerni kell a $P(X_j|Y_i)$ eloszlásának típusát, az eloszlás paramétereit pedig statisztikai módszerrel kell becsülni. Normális eloszlás esetén a várható értéket a mintaátlaggal, a szórásnégyzetet pedig a korrigált empirikus szórásnégyzettel lehet becsülni. Ekkor a keresett valószínűség:

$$P(X_j = x_j|Y_i) = \frac{1}{s_{ij}^* \sqrt{2\pi}} e^{\frac{(x_j - X_{ij})^2}{2{s_{ij}^*}^2}} \tag{18}$$

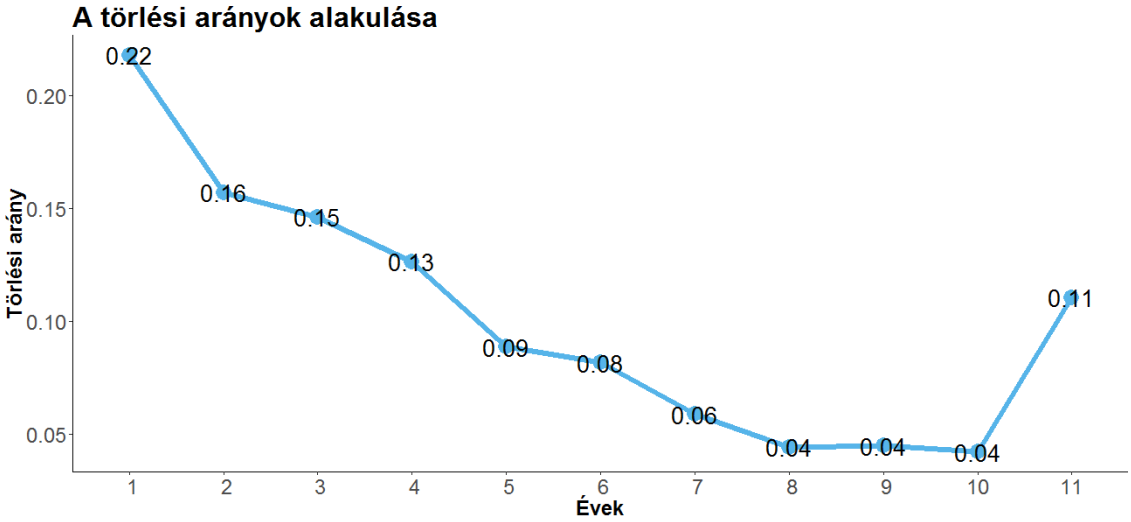
(Barber, 2012)

4. Az elemzéshez használt adatbázis bemutatása

Az elemzésemhez használt adatok egy külföldi biztosítótó egyik nyugdíjbiztosítási állományát tartalmazza 2017.október 20-ai állapot szerint. Első lépésként az adatokat minden személyes információtól semlegessé tettem, úgy, hogy megfeleljen a GDPR[11] követelményeinek. Ez nemcsak azért fontos, hogy ügyfelek személyes adatai biztonságban legyenek, de néhol az elemzéshez is szükségesek voltak. Például a születési dátum átkonvertálása, életkor változóvá.

A biztosítók az adatokat legtöbbször nem egy helyen, hanem különböző adattáblákban tárolják, annak megfelelően különválasztva, hogy milyen információt tárolnak. Ez ebben az esetben is így volt. Három különböző adattábla állt a rendelkezésemre: egy tartalmazta az ügyfelek, egy a szerződések és egy a szerződésekhez tartozó ügynökök adatait. Az elemzés megkezdéséhez, az adatok egyesítését kellett megoldanom, amelyhez egy úgynevezett kulcsra van szükség. Szintén bevett gyakorlat a biztosítóknál, hogy ez a kulcs a *szerződésszám*. Az így létrejött táblázat, mely 506280 sorból és 21 oszlopból áll, már alkalmas lehet az elemzés elkészítésére. A törléseknek más hatása van a biztosító eredményére, ha az első évben vagy a tízedik évben következik be, ezért a meglévő adatbázist három külön csoportra bontottam és ezen a három csoporton végeztem el az elemzést. A csoportok kialakítása a következők alapján történt. Az első adattáblában azok a szerződések vannak, amiket a szerződéskötéstől számítva legalább 365 napon keresztül kerültek megfigyelésre. A második csoportba azokat a megfigyeléseket gyűjtöttem, akiket legalább öt évig, míg a harmadik csoportban lévő adatokat tíz évig lehetett megfigyelni. Azért ezeket az időpontokat választottam, mert ez a három legfontosabb időpont egy szerződés tartama során, hiszen az ötödik év végétől lehetséges a szerződések visszavásárlása (40%-os visszavásárlási büntetéssel), míg a tízedik év után visszavásárlási büntetés nélkül. Az első évben bekövetkező törlések aránya szinte mindig a legmagasabb egy állományon.

[11] General Data Protection Regulation: 2018.május 25-től életbelépő rendelet. Az EU-n belüli adatkezelést szabályozza.



2. ábra: Törlési arányok a kockázatviselés kezdete utáni években

A 2. ábran látható, hogy a törlési arány tényleg az első évben a legmagasabb, utána fokozatosan csökken, majd a tizedik év után újra megugrik. Érdekes jelenség, hogy a második, harmadik és negyedik évben is viszonylag sokan törölnek, pedig ekkor a szerződés szerint semmilyen összeg nem jár vissza a szerződőnek. Ötödik évtől viszont a törlés már visszavásárlást jelöl.

4.1 Változók bemutatása

A modellezéshez használt kezdeti változók minden esetben megegyeznek, csak a megfigyelések számában különböznek. Ebben a fejezetben a változók jelentését és néhány leíró statisztikát mutatok be.

Változó név	Jelentés
age	Életkor
AgeCat	Életkor kategorizálva
agency_region	Ügynök területi egysége
ape	Éves díj
ApeCat	Éves díj kategorizálva
contract_category	Szerződés típusa
contract_id	Szerződésszám
contract_term	Szerződés tartama
contract_term_Cat	Szerződés tartama kategorizálva
contract_terminated_x	Törölték-e a szerződést
contract_theoritical_term_year	Biztosítás tartama
contract_theoritical_term_year_Cat	Biztosítás tartama kategorizálva
distribution_channel	Értékesítési csatorna
education	Iskolai végzettség
gender	Nem
marital_state	Családi állapot
occupation	Foglalkozás
payer_is_client	A biztosított a díjfizető
payment_frequency	Díjfizetési gyakoriság
payment_method	Díjfizetés módja
riskcommencing_year	Kockázatviselés kezdete

2. táblázat: A felhasznált adatbázis változói és jelentései

A 2. táblázatból látható, hogy az adatbázis nagyrészt kategorikus változókból áll, numerikus változóból csak az *age,* az *ape, contract_term* és a *contract_theoritical_term_year* elérhető. A változók közül a legtöbb egyértelmű, azonban néhány részletesebb kifejtést igényel.

agency_region: Többek között (Kim, 2005) is bizonyította, hogy makrogazdasági hatások is szerepet játszanak a törlések alakulásában. Ilyen adatok nem álltak rendelkezésemre, ezért erre a feladatra területváltozókat használtam, abban bízva, hogy egy ilyen típusú változó eredményesen tudja reprezentálni a gazdasági helyzetet. Ezt egy magyarországi példával könnyen bemutatható lenne, hiszen Szabolcs-Szatmár-Bereg

megyében rosszabb gazdasági körülmények vannak, mint Budapesten vagy Győr-Moson-Sopron megyében. Azonban a számomra rendelkezésre álló adatokban a szerződéshez tartozó területi változóban túl sok kategória volt. Magyarországi példán ilyen esetben a változókat összevonnám: a 20 megye átcsoportosítható 7 régióvá. Azonban ennél az adatbázisnál a GDPR miatt a tényleges területi változók helyett csak olyan adatok álltak rendelkezésemre, amelyek nem „érzékenyek" („megye1", „megye2", stb), így az ilyen összevonásra nem volt lehetőségem. Ami maradt, az az ügynök területi csoportosítása, ami a 11 kategóriájával már jobban kezelhető, mint a szerződő területi besorolása. Így azzal a feltételezéssel, hogy az ügynök azon a területi egységen dolgozik, ahová be van jelentve, ez már eleget tesz az eredeti célkitűzésemnek.

education: Az eredeti adatbázisban összesen 10 féle kategóriája volt ennek a változónak, amiről úgy ítéltem meg, hogy túl részletes megbontás, ezért új kategóriákat alkottam: Alapfokú (*Elementary*), középfokú (*UnderGraduate*) és felsőfokú (*Graduate*). Az eredeti és az új értékek megtalálhatóak a függelék 16. táblázatában.

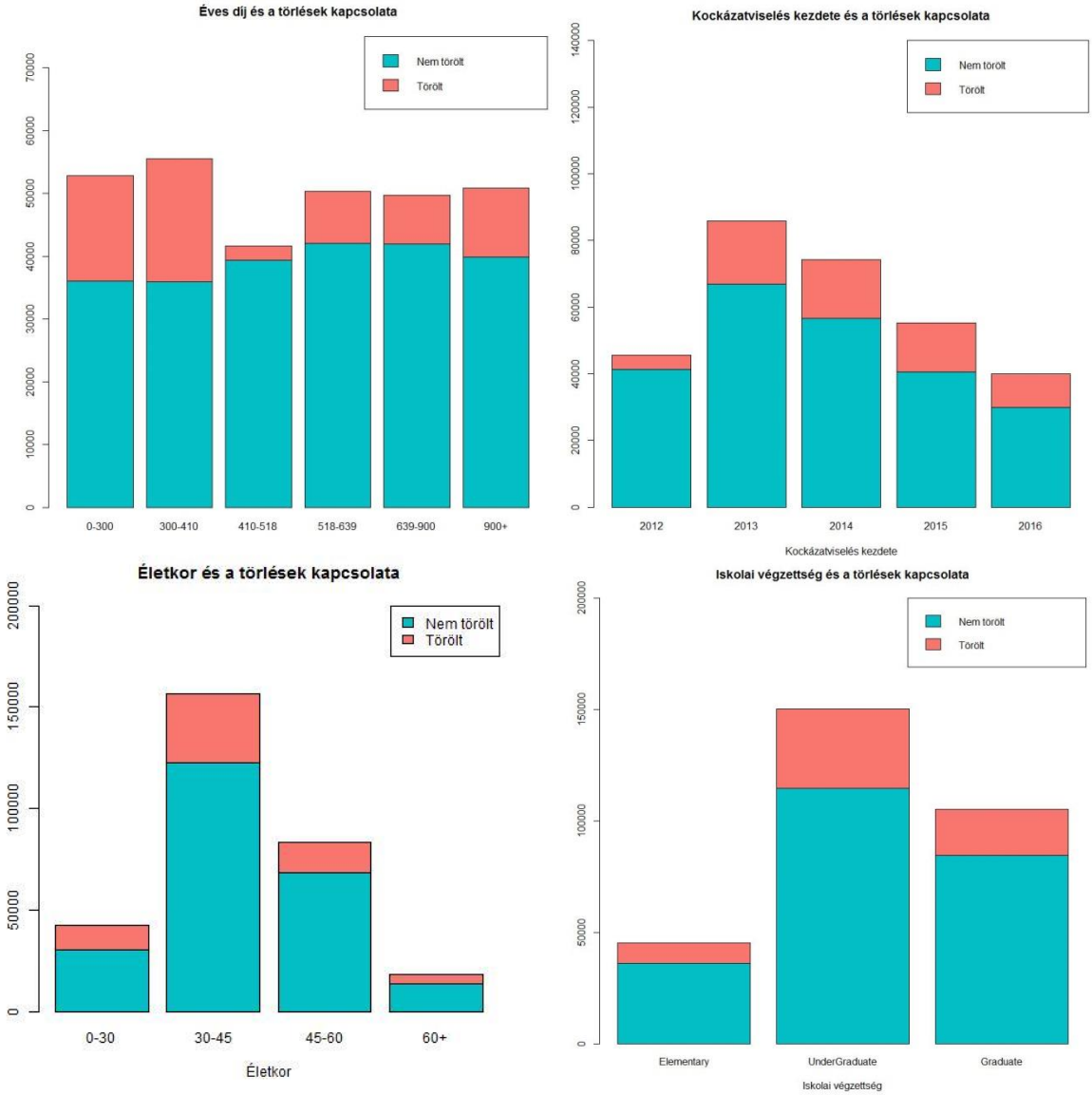
occupation: Hasonlóan az *education* változóhoz, ennél is túl részletes volt a megbontás. Nincs értelme külön törlési modellt építeni például a rendőrökre és csendőrökre, ezért ezekből is új kategóriákat hoztam létre. A szokásos fizikai (*blue collar)* és szellemi munkás (*white collar*) kategóriákba nem tudtam volna az összes változót besorolni, ezért felhasználva (Basu, 2015)-ös cikkét létrehoztam egy úgynevezett *pink-collar* kategóriát, ebbe tömörítve azokat a foglalkozásokat, amelyek hagyományosan nők végeznek (például: ápolónő, fodrász vagy háziasszony). Amiket még a fenti három kategóriába se tudtam besorolni (például: vállalkozó vagy diák) az *egyéb*be soroltam.

Éves díj: Az elemzések során, ahol a modell engedte kategorizált változót használtam, mert az eredeti változó szórása nagyon magas volt. A csoportok határait úgy próbáltam meg beállítani, hogy a minta elemszámok közel legyenek egymáshoz. Az értékek euróban értendőek.

4.1.1 Contracts01

A contracts01 nevű tábla tartalmazza azokat a szerződéseket, amik legalább egy évvel a cenzorálást megelőzően bekerültek a rendszerbe, összesen 300895 megfigyelést. Az eredeti megfigyelésszám körülbelül 200000-el csökkent. Ez vagy annak köszönhető, hogy a kapott adatbázis nem teljes, vagy valamilyen adatrögzítési hiba történt, ugyanis

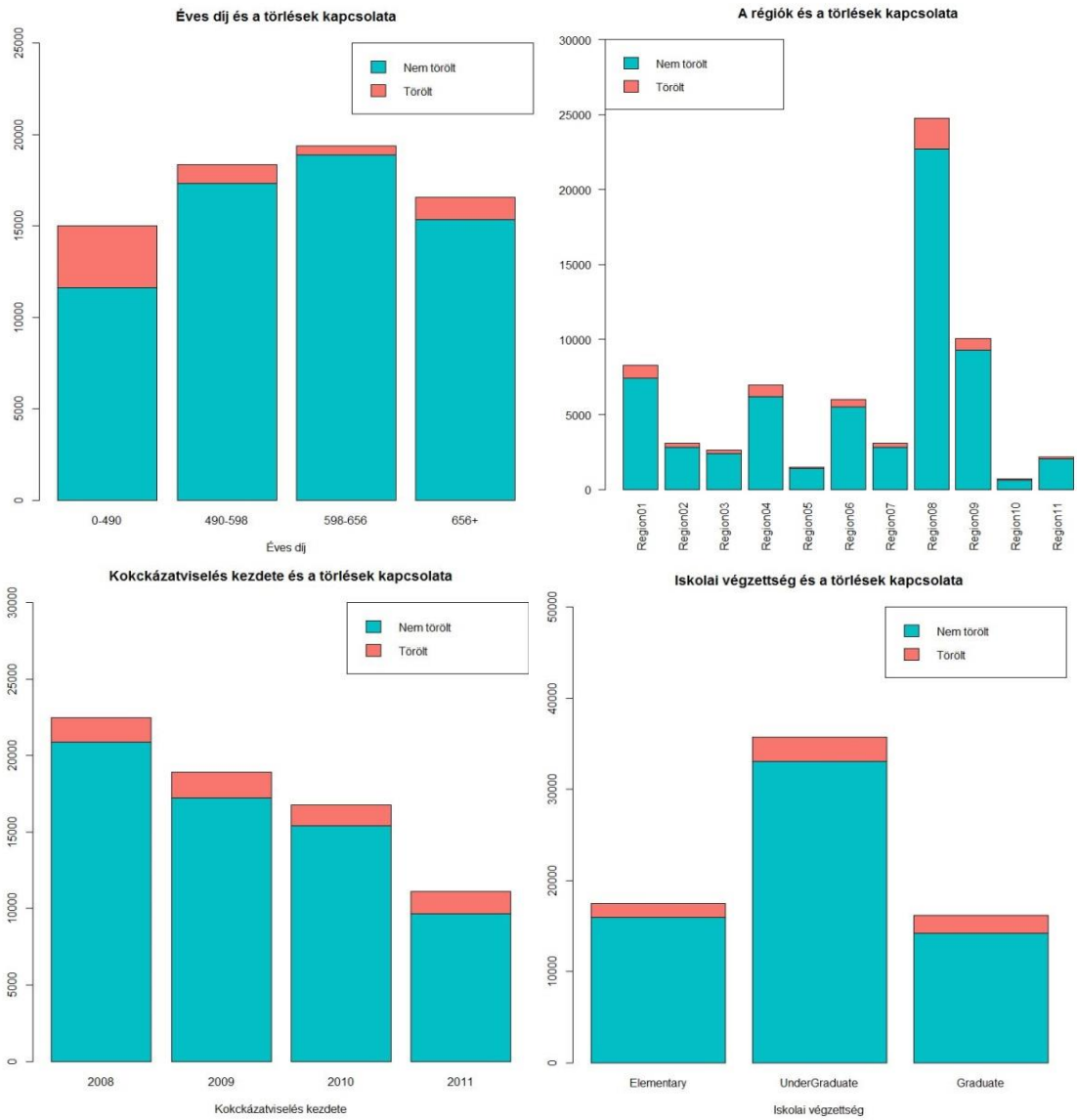
míg az első megfigyelések már 2004-től rögzítésre kerültek, az első törlés csak 2013-ban lett regisztrálva. A való életben ennek a problémának utána lehetne járni, nekem erre nem volt lehetőségem, ezért a problémát úgy oldottam meg, hogy a contracts01 táblába csak azokat a megfigyeléseket gyűjtöttem, amelyeknél a kockázatviselés kezdete 2012. január 1. utáni. Kategorikus változók bemutatására egy egyszerű módszer az oszlopdiagramok használata. Alább bemutatom néhány fontosabb változó oszlopdiagramját. Ezekről leolvasható a változó lehetséges értéke, az egyes kategóriák előfordulási gyakorisága, valamint a törölt és nem törölt szerződések aránya.



3. ábra: *az első éves törlések vizsgálatához használt változók közül néhány*

4.1.2 Contracts06

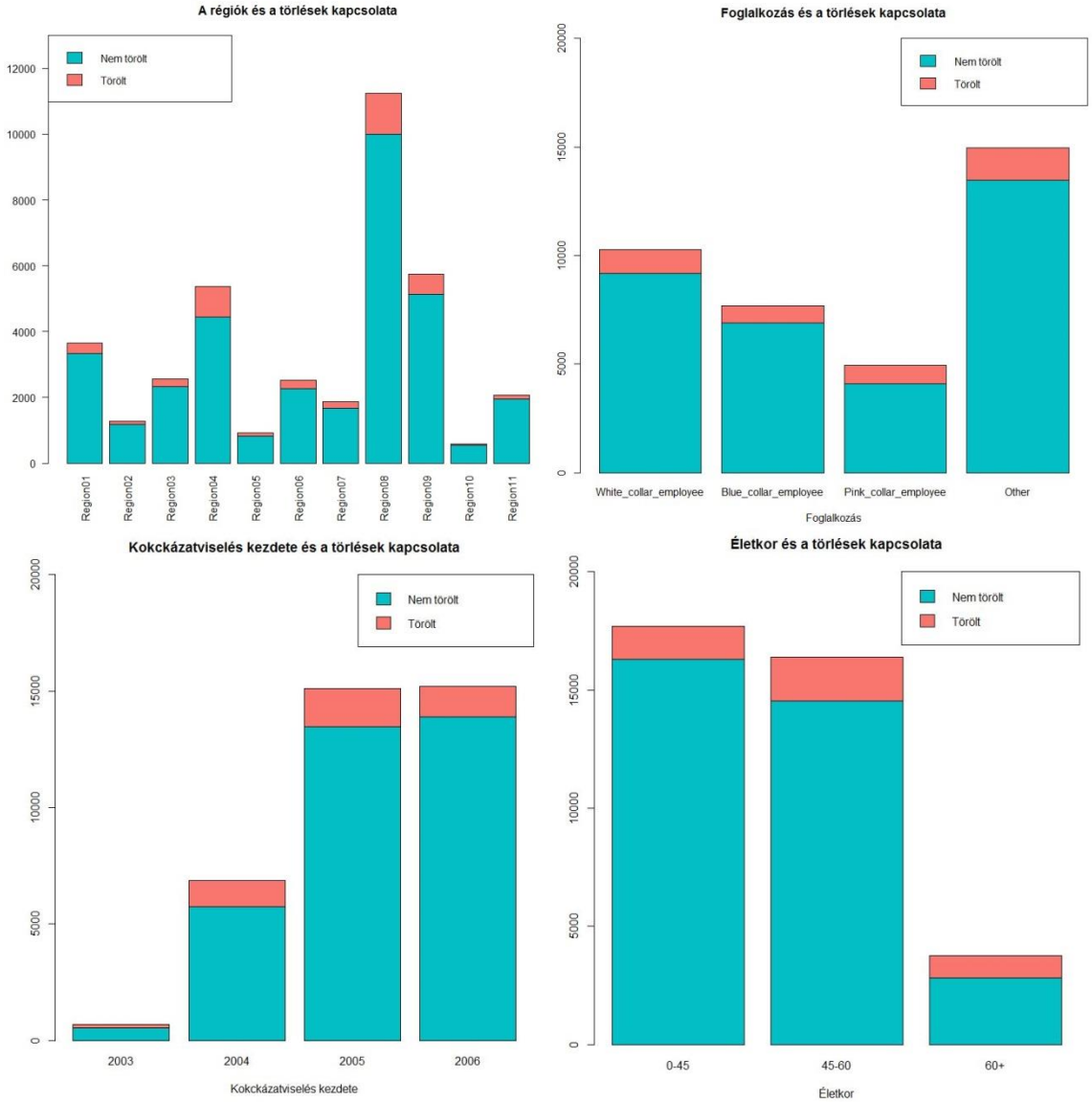
A contracts06 nevű tábla tartalmazza azokat a szerződéseket, amelyek legalább hat évvel a cenzorálást megelőzően bekerültek a rendszerbe, de az első öt évben nem töröltek, így elkészíthetőek a törlési modellek az ötödik évre. Technikailag egy szerződés akkor kapott „törölt” státuszt, ha tartama kisebb, mint hat év. Összesen 69305 megfigyelés tartozik ide.



4. ábra: az ötödik éves törlések vizsgálatához használt változók közül néhány

4.1.3 Contracts11

A contracts11 az előzőekhez nagyon hasonlóan készült, ezt a táblát használtam a tíz éves törlések vizsgálatához, összesen 37849 megfigyelés tartozik ide.



5. ábra: a tizedik éves törlések vizsgálatához használt változók közül néhány

5. A bemutatott módszerek alkalmazása

Ebben a fejezetben a már ismertetett adatbázisokon, miként teljesítettek a korábban bemutatott adatok. Az elemzéshez R programot használtam.

5.1 Logisztikus regresszió

A logisztikus regresszió kevéssé hajlamos a túlillesztésre, ettől függetlenül itt is célszerű tanuló és tesztelő adatbázisra szétválasztani az adatokat, hiszen az ilyen elemzések célja az, hogy az eredményként kapott modell képes legyen bármilyen új adatpontot a lehető legpontosabban kategorizálni. A tanuló adatbázis 80%-a teljes mintának, míg a fennmaradó 20%-on teszteltem le a modell pontosságát.

5.1.1 Első éves törlések

A modellezést az összes változó bevonásával kezdtem, majd a legkevésbé szignifikáns változót eltávolítottam és újrafuttattam a modellt. Ezt addig ismételtem, míg csak szignifikáns változó maradt, melynek paraméterei megtalálhatóak a függelékben, a 17. táblázatban. Ebben megtalálhatóak a magyarázó változókra vonatkozó becslések, valamint azok szignifikanciaszintje is. Megfigyelhető, hogy a kategorikus változók közül egy mindig kimarad a felsorolásból. Ez azért van, mert az egyes kategóriákhoz becsült paraméterek értéke a kimaradt változót, mint referenciaszintet használja. Például a kategorizált éves díj (*ApeCat*) változó esetén ez a referencia kategória a 0-300 euró közötti sávba esik. A kapott eredményeket úgy lehet értelmezni, hogy ha egy szerződés éves díja 300-410 euró között van, akkor annak az odds-a, hogy az ügyfél törölni fog exp(0.11635)-szerese annak, mintha 0 és 300 euró között lenne. Hasonlóan a többi változónál. Látható tehát, hogy a paraméterek értelmezése nehezebb, mint a lineáris regresszió esetében, de egy-egy változó hatása könnyen eldönthető a predikció előjelének vizsgálatával. Ha pozitív, akkor növeli az oddsot, ha negatív, akkor csökkenti. A logisztikus regresszió egyik feltételezése, hogy a változók között nem áll fenn multikollinearitás. Ennek a tesztelésére a VIF (variancia infláló faktor) mutató használható, mely 1-es értéke teljes korrelálatlanságot jelent a változók között.

Elméletileg a VIF végtelen nagy is lehet, de hüvelykujjszabálynak az 5-ös értéket szokták használni a multikollinearitás jelenlétére. A VIF értékeket tartalmazza a függelékben lévő 18. táblázat. Ebből jól látható, hogy nincs jele multikollinearitásnak, így a magyarázóváltozók között nincs lineáris kapcsolat, tehát nincs akadálya annak, hogy a modellt a tesztelő állományon alkalmazzam. A prediktáláshoz mindig szükség van egy úgynevezett cut value-ra, ami az az érték, ami felett az előrejelzés értéke 1-es lesz, alatta pedig 0. Ebben az esetben ez azt jelenti, hogy ha modell által prediktált valószínűség 0,25 felett lesz, akkor törlést jelez előre, ha alatta, akkor a „nem törlést" jelzi. Az előrejelzés pontossága keresztvalidációval ellenőrizhető.

	Valós állapot	
Klasszifikáció	0	1
0	32782	5041
1	14262	8093

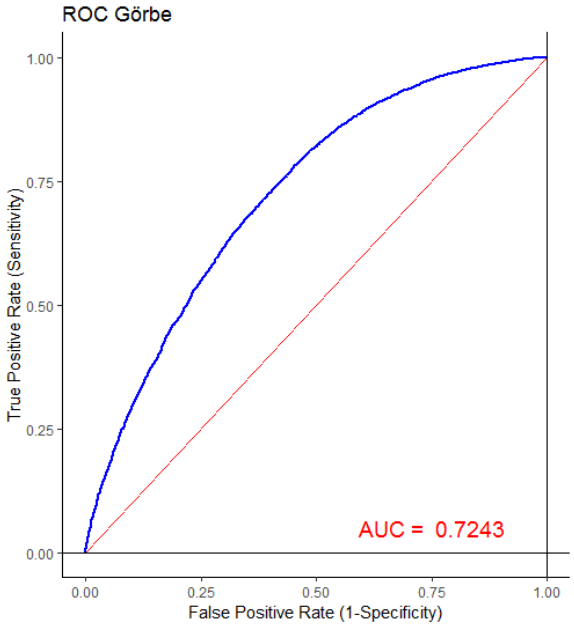
3. táblázat: A logisztikus regresszió elsőéves törlésekhez kötődő konfúziós mátrixa

A konfúziós mátrix diagonálisában találhatóak az úgynevezett *true postive* és *true negative* értékek. Ezek azok a megfigyelések, amelyeket a modell helyesen klasszifikált, jelen esetben ez 67,92%. Ezen kívül a mátrixban megtalálható az első és másodfajú hiba[12] értéke is. Elsőfajú hibát 5041-szer, másodfajú hibát 14262-szer követett el a modell. A konfúziós mátrix kicsi, de valójában sok információt hordoz magában, azonban mégsem ez a legjobb módja a modell pontosságának vizsgálatára, mert a klasszifikáció pontosságát nagymértékben befolyásolja a cut-value értéke. A modell kiértékelésének egy másik lehetséges mutatószáma az AUC *(area under curve)* érték, ami a ROC görbe alatti terület. A ROC görbe egy monoton növő függvény a $[0,1] \times [0,1]$ négyzetben, amelyben minden lehetséges cut-value esetén ábrázolja az úgynevezett *true positive* és a *false postive* arányokat. A $(0,1)$ pont érintése esetén a modell teljesen tökéletes, valamint véletlenszerű a klasszifikáció, ha a ROC görbe egybeesik az $y = x$ egyenessel.

Az első éves törlésekhez készült logisztikus regressziós modell látható a 6. ábrán. Általánosságban hüvelykujjszabálynak a 0,7-es küszöbérték számít elfogadható modellnek. Jelen esetben a görbe alatti terület 0,7243.

[12] Első fajú hiba *(false positive)*: Első fajú hibának nevezzük azt, amikor egy megfigyelést a modell törölt szerződésnek azonosítja, de valójában túlélő

Másodfajú hiba: *(false negative)*: Másodfajú hibának nevezzük azt, amikor egy megfigyelést a modell túlélőnek feltételez, de valójában töröl



6. ábra: Az első évben bekövetkező törlésekhez készült logit modell ROC görbéje

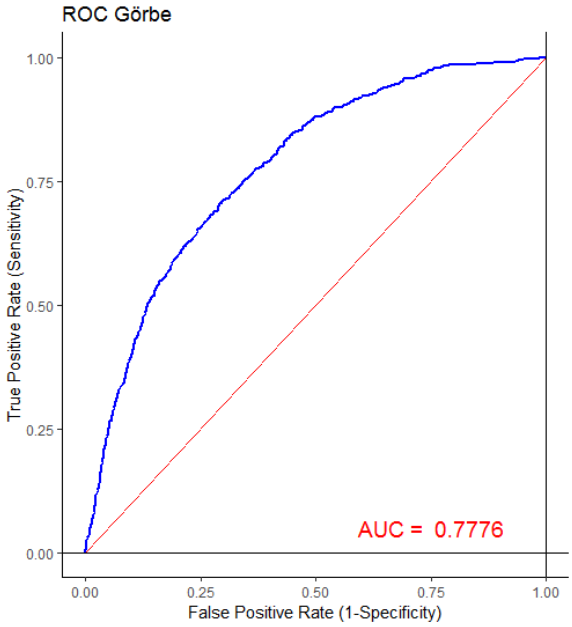
5.1.2 Ötödik éves törlések

Az előző modellhez hasonlóan, itt is 80-20 arányban bontottam meg az adatbázist tanuló és tesztelő állományra, és a modellszelekciót változónként végeztem el. Azonban ennél a modellnél kevesebb szignifikáns változó lett a végső modellben. Valószínűleg ez azért van, mert a szerződések csupán 9%-a törölt, így nem minden kategória alapján lehet megfelelően szétválasztani a megfigyeléseket. A modell által becsült koefficiensek a függelék 19. táblázatában találhatóak. Az előző modellhez hasonlóan a VIF értékei itt is minden változó esetén öt alatt van (20. táblázat). A konfúziós mátrix elkészítése 10 százalékos cut-valueval történt, igazodva a mintában lévő 9 százalékos törlési arányhoz.

	Valós állapot	
Klasszifikáció	0	1
0	9172	391
1	3460	837

4. táblázat: A logisztikus regresszió ötödik éves törlésekhez kötődő konfúziós mátrixa

A kevesebb megfigyelés ellenére ez a modell pontosabb klasszifikációra képes, mint az első, az adatok 72,22%-át becsülte pontosan. Ugyanezt támasztja alá az AUC mutató is a 0,7776-os értékével.



7. ábra: Az ötödik évben bekövetkező törlésekhez készült logit modell ROC görbéje

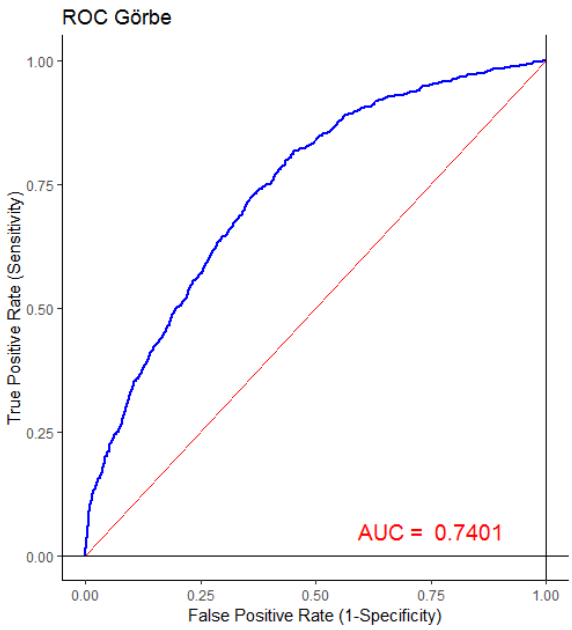
5.1.3 Tízedik éves törlések

A modellezés lépései megegyeznek az előző két modellnél bemutatottokkal. A koefficiensek a függelék 21. táblázatában találhatóak. A VIF értékei szintén a függelékben, a 22. táblázatban találhatóak. A konfúziós mátrix (igazodva a mintához), ennél a modellnél is 0,1-es cut-valuval készült.

	Valós állapot	
Klasszifikáció	0	1
0	4391	249
1	2338	591

5. táblázat: A logisztikus regresszió tízedik éves törlésekhez kötődő konfúziós mátrixa

A mátrixból kiolvasható helyes klasszifikációs arány 65,82%-os, ami az eddigi modellek közül a legrosszabbnak mondható, azonban a ROC görbe alatti terület alapján ez a modell pontosabb, mint az első.



8. ábra: A tizedik évben bekövetkező törlésekhez készült logit modell ROC görbéje

A logisztikus regresszió, egy hagyományos statisztikai módszertan, melyet már hosszú ideje alkalmaznak a matematikusok, statisztikusok. A továbbiakban azt fogom vizsgálni, hogy esetleges melyik modellel lenne célszerű helyettesíteni jobb eredmény érdekében.

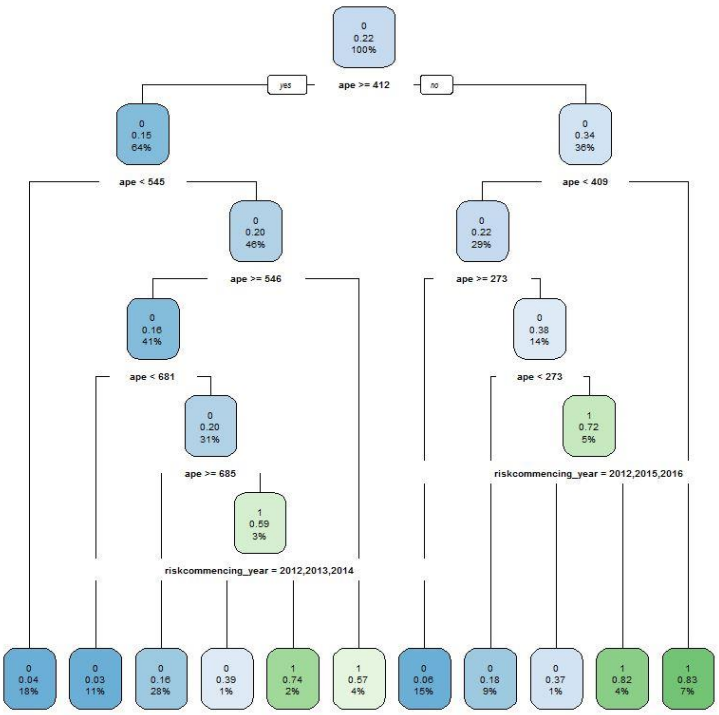
5.2 Döntési fa

A döntési fák egyik jó tulajdonsága, hogy bármilyen adaton használható, ezért az első modellbe minden változót beletettem, és engedtem, hogy a fa felépítse magát. A fa növekedését általában a *minsplit*, a *minbucket* és a *cp* paraméterrel szokták szabályozni. A *minsplit* és a minbucket egymástól függő paraméterek, előbbi az egy ágon lévő megfigyelés számot határozza meg, míg utóbbi ugyanezt, csak egy levélen. Alapesetben a paraméterezés úgy történik, hogy egy levélen minimum harmadannyi megfigyelés legyen, mint egy ágon, ahhoz, hogy a vágás megtörténjen. A *complexity parameternek (cp)* két fontos szerepe van. Egyrészt meggyorsítsa az eljárást, ugyanis, ha egy vágás nem javítja a modellt legalább „cp”-vel, akkor az a vágás nem is jön létre, ezzel jelentős számítási igényt spórolva. Másrészt hasonló elvek alapján gátolja a túlillesztést is.

A döntési fák egyik legfontosabb jellemzője, hogy hajlamosak a túlillesztésre, ezért a logisztikus regresszióhoz hasonlóan a modellt a 80%-is tanuló állományon készítettem el, majd a 20%-os tesztelő állományon megvizsgáltam a klasszifikáció pontosságát.

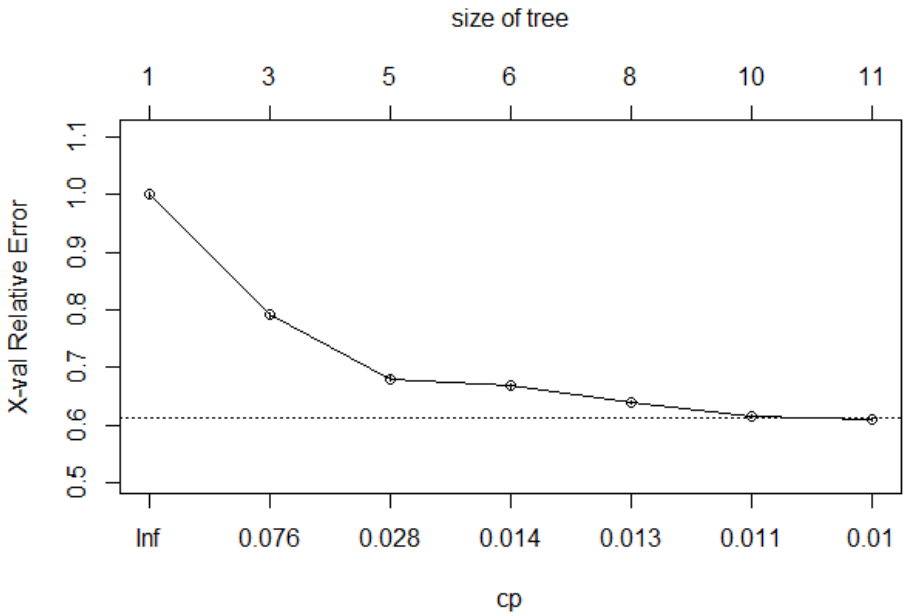
5.2.1 Első éves törlések

A fent említett paraméterek közül sok lehetőséget kipróbáltam, de a legjobb modell az alapbeállításokkal született, amely a 9. Ábraán található. Látható, hogy ez egy hat mélységű fa, ugyanakkor az értelmezése még mindig nem okoz gondot, főleg azért, mert csak az éves díj és a kockázatviselés kezdete változókat használja. Egy levél értelmezése úgy történik, hogy a fát fentről lefelé haladva vizsgáljuk, mígnem a kívánt levélről minden információ ismerté válik. A bal alsó levél például azokat a megfigyeléseket tartalmazza, amelyekben az éves díj 412 és 545 között van. Ide tartozik a megfigyelések 18%-a, amelyeknek a törlési valószínűsége 4%.



9. Ábra: Az első éves törlések modellezésére készített döntési fa

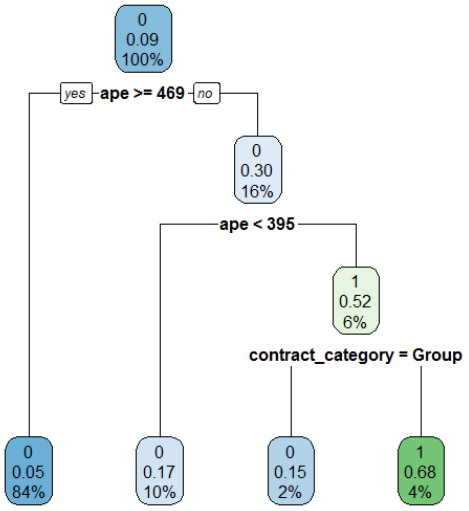
A keresztvalidációt megkönnyíti a *printcp()* parancs, amely a keresztvalidációs hibát mutatja meg különböző *cp* paraméterek mellett. A *plotcp()* parancs pedig ezeket az értékeket ábrázolja, így segítve a döntést. A 10. Ábraán látható, hogy a hiba akkor a legkisebb, ha *complexity parameter* értéke 0,1. A klasszifikáció 86,56%-ban sikeres.



10. Ábra: A relatív hiba nagysága az első éves törléseknél

5.2.2 Ötödik éves törlések

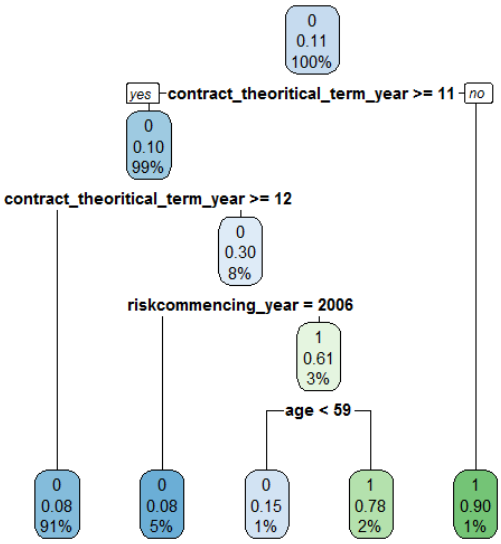
A legkisebb hibát most is az alapbeállítással készült fa adja, azonban ebben az esetben ez egy sokkal egyszerűbb fa, mint az előző. Ennek ellenére a tesztelő állományon 92,86%-ban helyesen klasszifikál. Az éves díj itt is fontos változó, valamint egy új változó alapján is megbontja az adatokat a döntési fa, ez pedig a *contract_category*. A „Group” érték arra utal, hogy a szerződést csoportban kötötték. Az ábráról megállapítható, hogy azokon a szerződéseken a legvalószínűbb a törlés, amelyiknél az éves díj 395 és 468 között van, valamint nem csoportos biztosítás. Ekkor a törlés valószínűsége 68%.



11. Ábra Az ötödik éves törlések modellezésére készített döntési fa

5.2.3 Tízedik éves törlések

A harmadik döntési fa szintén különbözik az előző kettőtől, ezzel is bizonyítva a modell sokszínűségét. Végeredményben egy négy mélységű fát kaptam, mely az éves díjat nem tartalmazza. Ennek egyik oka az lehet, hogy azon ügyfelek, akik tíz évig tudták fizetni a díjat, már nem annyira érzékenyek a biztosítási díjra. A fontosabb változók ebben az esetben az életkor, és a biztosítás tartama. A tesztelő adatokat 91,39%-os pontossággal klasszifikálja az alábbi döntési fa:



12. Ábra: A tizedik éves törlések modellezésére készített döntési fa

5.3 Véletlen erdő

Már a döntési fákkal is sokkal pontosabb eredményt sikerült elérni a logisztikus regressziónál, ezt szerettem volna javítani a véletlen erdő algoritmussal. Ez a módszer lényegében sok döntési fa – jelen esetben 1000 – elkészítését jelenti, néhány plusz paraméter bevonásával. Az egyik ilyen az *mtry* paraméter mellyel azt lehet beállítani, hogy egy-egy potenciális vágásnál hány változót vizsgáljon. Azért célszerű vágásonként néhány változót kihagyni, mert ha van egy nagyon erős változó (az első éves törlések vizsgálatánál látható volt, hogy az *ape* egy ilyen), akkor az minden fában benne lesz, így eltorzítva a végeredményt. A véletlen erdők futtatásakor az mtry paraméter értékeit 2 és 12 között változtattam, végső paraméternek azt választva, amelyik erdő a legjobban teljesített a teszt állományon. Az így kapott konfúziós mátrixok:

	Viszonyítási alap	
Pred	0	1
0	45277	3036
1	1767	10098

	Viszonyítási alap	
Pred	0	1
0	12429	813
1	203	415

	Viszonyítási alap	
Pred	0	1
0	6701	623
1	28	217

6. táblázat: Az egy, öt és tíz éves törlésekre vonatkozó véletlen erdő modellek konfúziós mátrixai

A 6. táblázatból könnyen kiszámolható a klasszifikáció pontossága: 92,02%; 92,67%; 91,40%. Az első erdőnél hat, a másodiknál öt, míg a harmadiknál kettő változót vizsgáltam vágásonként, mert ezekkel lehetett a legpontosabb eredményeket elérni.

5.4 k-NN

A legközelebbi szomszéd módszere csak numerikus prediktorokat tud kezelni, így a modellbe csak két magyarázó változó került be (normalizálás után), az életkor és az éves díj. Létezik egy hüvelykujjszabály *k* értékére, ez a minta elemszámának négyzetgyöke. Ez a legkisebb adatbázison is körülbelül *k* = 200-at jelentene. Ekkora érték választása óriási számolási igényt igényel, így én csupán 1-től 100-ig számoltam ki a klasszifikációs hibákat és azok közül választottam a legjobbakat.

k=1	Viszonyítási alap	
Pred	0	1
0	40846	6282
1	5575	7476

k=20	Viszonyítási alap	
Pred	0	1
0	12609	41
1	1179	32

k=49	Viszonyítási alap	
Pred	0	1
0	6690	52
1	662	166

7. táblázat: k-NN modellek konfúziós mátrixai

A helyes találati százalékok rendre: 80,3%; 91,2%, 90,57%. A magas százalékok szintén annak köszönhetőek, hogy a mintában alacsony a törlésszám, így ezeknek a sikertelen klasszifikálása, nem okoz akkora pontosságvesztést.

5.5 Naív Bayes

A naív Bayes algoritmus eredményeképp megkaphatóak a kategóriánkénti becsült feltételes valószínűségeket. Példaként bemutatom az első éves törlések *éves díj* kategóriáira kapott outputot:

	ApeCat					
	0-300	**300-410**	**410-518**	**518-639**	**639-900**	**900+**
0	0,1532	0,1530	0,1671	0,1780	0,1789	0,1699
1	0,2557	0,3010	0,0355	0,1252	0,1180	0,1647

8. táblázat: ApeCat változóra vonatkozó feltételes valószínűségek a naív Bayes módszerrel

Annak a feltételes valószínűsége, hogy a szerződés éves díja 0 és 300 euró között van, feltéve, hogy a szerződés törölt az 0,1532, míg annak a feltételes valószínűsége, hogy a szerződés éves díja 0 és 300 euró között van, feltéve, hogy a szerződés **nem** törölt az 0,2557.

A naív Bayes algoritmus lefuttatása után a szokásos módon kiszámolhatóak a törlése kapott becslések:

	Viszonyítási alap	
Pred	0	1
0	68342	17593
1	2224	2108

	Viszonyítási alap	
Pred	0	1
0	18632	1723
1	317	119

	Viszonyítási alap	
Pred	0	1
0	9854	1160
1	240	100

9. táblázat: A naív bayes-i klasszifikáció eredményei

A modell klasszifikációs pontosság rendre: 78.05%; 90.19%; 87.67%. Százalékos értékkel nézve a második és harmadik modell egészen kiemelkedő, azonban az látható, hogy a törölt szerződéseket kevésbe tudja prediktálni, ami egy elég nagy kritika egy törlési modellel szemben.

5.6 SVM

A bemutatott módszerek közül a modern SVM a legújabb modell, és robusztussága miatt nagyon népszerű, így ettől vártam a legjobb eredményeket. A klasszifikáció gyakorlatban az úgynevezett tuning paraméterek beállításával automatikusan történik. Ezek a *gamma* és a *költség* paraméterek. Gamma határozza meg, hogy egy-egy pontnak mekkora hatása lehet a támaszvektorok kialakításában. Ha egy x_j támaszvektornak a gamma értéke kicsi, az azt jelzi, hogy annak a támaszvektornak nagy hatása van a többi x_i pont klasszifikálásában, még akkor is, ha a köztük lévő távolság nagy. Nagy értéke ennek az ellentettjét mutatja, tehát azt, hogy az adott pontnak nincs nagy szerepe a támaszvektorok kialakításában. A költség paraméter a (15)-ös képletben szereplő *C* érték, azt szabályozza, hogy hány megfigyelés kerülhet a hipersík rossz oldalára. Nagy *C* érték megengedőbb a határsértésekkel szemben, mint a kicsi.

A finomhangolás (tuning) lényege, hogy a megadott paraméterekkel elkészülő modellek közül már keresztvalidációval megállapítható, hogy melyik paraméterrel lehet a legjobb klasszifikációt elérni. Azonban ez a módszer meglehetősen számításigényes, a futtatás közben én is problémába ütköztem, melyet az adatbázis csökkentésével tudtam csak megoldani. Így az SVM modellt csak az első éves törlésekre alkalmaztam és azt is úgy, hogy véletlen mintát vettem az adatbázison. Az így meglévő 30000 modellponton futtattam le a finomhangolást, melynek kezdeti paramétereit 10^{-5} és 10^{3} között állítottam be.

Az eredmények a szokásos konfúziós mátrixban találhatóak, a klasszifikációs pontosság 77,67%

10. táblázat: Az SVM klasszifikáció eredményei

	Viszonyítási alap	
Pred	0	1
0	19754	6392
1	306	3548

6. Következtetések

Dolgozatomban arra törekedtem, hogy olyan módszereket ismertessek meg amellyel a biztosító állományára vonatkozó törlések pontosabban modellezhetők. Ehhez elengedhetetlen volt ismertetnem a törlési kockázat Szolvencia II-es definícióját, a benne rejlő veszélyeket és az előrejelzés pontosságának fontosságát. A törléseket viselkedési közgazdaságtani szempontból is elemeztem, melyet azért tartottam relevánsnak, mert ez is egy új képet adhat a törlések alakulásáról. Arra próbáltam felhívni a figyelmet, hogy nem elég pusztán matematikai modelleket készíteni, ahhoz, hogy pontos képet kapjon a biztosító az állományán bekövetkező törlésekről. Érdekes kérdés lehet, hogy hogyan lehet a meglévő modellekbe beleépíteni ezen viselkedési közgazdaságtani szempontokat.

Az elméleti bevezető után bemutattam az alkalmazott gépi tanulási és adatbányász módszerek alapfeltevéseit és matematikáját, melynek értése elengedhetetlen a modellezés pontos kivitelezéséhez.

A modellezéshez használt kezdő adatbázis egy 506280×21-es mátrixból állt. Az adatbázis mérete miatt azonban volt lehetőségem több különböző időpontra is törlési modellt készíteni. Ezt azért tartottam fontosnak, mert feltételeztem, hogy a törlést más befolyásolja a szerződés első évében, mint például a tartam közepén. Éppen ezért három időpontot választottam: a szerződés első, ötödik és tizedik évi törléseit vizsgálva. A modellezést két dolog nehezítette, az adatok mennyisége nagyon nagy számításigényt követelt meg, amelyhez nem is mindig volt egy elég hagyományos számítógép. A biztosítóknak azonban ez nem jelenthet gondot a 21. században, amikor könnyen, gyorsan és olcsón juthatnak hozzá plusz teljesítményhez. A másik dolog, ami nehézséget jelentett az alacsony törlésszám. Ez a biztosítónak jó hír, de a klasszifikációt nehezíti, ha kevés az előre jelzendő megfigyelés. A modellek teljesítményét a logisztikus regresszióval összehasonlítva vizsgáltam, mert a logisztikus regresszió a legrégebb óta használt és legelterjedtebb klasszifikációs módszer.

	1.év	5.év	10.év
Logisztikus regresszió	67,92%	72,22%	65,82%
Döntési fa	86,56%	92,86%	91,39%
Véletlen erdő	92,02%	92,67%	91,40%
k-NN	80,30%	91,20%	90,57%
Naív Bayes	78,05%	90,19%	87,67%

11. táblázat: Összesítő táblázat a felhasznált modellek klasszifikációs eredményeivel

A 11. táblázatból látható, hogy a logisztikus regressziónál mindegyik modell jobban teljesített, azonban ezeket az eredményeket fenntartással kell kezelni, mert a magas klasszifikációs eredmény annak köszönhető, hogy a mintában alacsony volt a törlésszám, így a másodfajú hiba értéke is alacsony. Ami azt jelenti, hogy pont azokat az ügyfeleket klasszifikálják a modellek tévesen, akikre valójában fókuszálnia kellene. A modellek további szofisztikálása, esetleges új változók létrehozása és bevonása, néhány modell paraméterének pontosabb meghatározása megoldást nyújthat erre problémára. Azonban minden ilyen változás nagyban függ a biztosító saját állományától, így ezen lépéseket már a dolgozat esteleges későbbi használójára bízom.

7. Irodalomjegyzék

Bacinello, A. R., 2003. Fair valuation of a guaranteed life insurance participating contract embedding a surrender option. *Journal of Risk and Insurance,* pp. 461-487.

Banyár, J., 2016. *Életbiztosítás.* Budapest: Budapesti Corvinus Egyetem.

Barber, D., 2012. *Bayesian Reasoning and Machine Learning.* 1st szerk. London: Cambridge University Press.

Basu, S., 2015. Health and pink-collar work. *Occupational Medicine* , pp. 2-6.

Bodon, F., 2010. *Adatbányaszati algoritmusok*.

Burkov, A., 2019. *The Hundred-page Machine Learning Book.*

Calí, C. & Lombardi, M., 2015. Some mathematical properties of the ROC curve and their applications. *Ricerche di Matematica,* pp. 391-402.

Campbell, J. és mtsai., 2014. *Modeling of Policyholder Behavior for Life Insurance and Annuity products,* hely nélk.: Society of Actuaries.

Cover, T. M., 1965. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers,* EC-14(3), pp. 326-334.

Cox, J. C., Ross, S. A. & Rubinstein, M. E., 1979. Option pricing: A simplified approach. *Journal of Financial Economics,* pp. 229-263.

Dataflair Team, 2017. *https://data-flair.training.* [Online]
Available at: https://data-flair.training/blogs/applications-of-svm/

EIOPA, 2011. *EIOPA Report on the fifth Quantitative Impact Study (QIS5) for Solvency II*

Eling, M. & Kiesenbauer, D., 2013. What Policy Features Determine Life Insurance Lapse? An analysis of the German market. *Working papers on risk management and insurance,* 81(2), pp. 241-269.

EU, 2009. *Az Európai Parlament és a Tanács 2009/138/EK irányelve (2009. november 25.) a biztosítási és viszontbiztosítási üzleti tevékenység megkezdéséről és gyakorlásáról (Szolvencia II)*

Fedoria, M. & Förstemann, T., 2015. *Lethal Lapses: How a Positive Interest Rate Shock Might Stress German Life Insurers,* Frankfurt am Main: Deutsche Bundesbank.

Grosen, A. & Lochte Jorgensen, P., 2000. Fair valuation of life insurance liabilities: The impact of interest rate guarantees, surrender options, and bonus policies.. *Insurance: Mathematics and Economics*, pp. 37-57.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2017. *An Introduction to Statistical Learning with Applications in R.* 8 szerk. New York: Springer.

Kim, C., 2005. Modeling Surrender and lapse rates with economic variables. *North American Actuarial Journal*, pp. 56-70.

Kuo, W., Chenghsien, T. & Wei-Kuang, C., 2003. An empirical study on the lapse rate: The cointegration approach. *Journal of Risk and Insurance*, pp. 489-508.

Patel, N. & Upadhyay, S., 2012. Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA. *International Journal of Computer Applications*, 60(12), pp. 20-25.

Shalev-Shwartz, S. & Ben-David, S., 2014. *Understanding Machine Learning from Theory to Algorithms.* New York: Cambridge University Press.

Vékás, P., 2016. *Az élettartam-kockázat modellezése.* Budapest

9. Függelék

	Halandóság	Hosszú élet	Rokkantság és betegség	Törlés	Költségek	Revízió	Katasztrófa
Halandóság	1	-0,25	0,25	0	0,25	0	0,25
Hosszú élet	-0,25	1	0	0,25	0,25	0,25	0
Rokkantság	0,25	0	1	0	0,5	0	0,25
Törlés	0	0,25	0	1	0,5	0	0,25
Költség	0,25	0,25	0,5	0,5	1	0,5	0,25
Revízió	0	0,25	0	0	0,5	1	0
Katasztrófa	0,25	0	0,25	0,25	0,25	0	1

12. táblázat: A Szolvencia II kockázati almoduljai közötti korrelációs mátrix

	Piaci	Partner	Élet	Egészség	Nem-élet
Piaci	1	0,25	0,25	0,25	0,25
Partner	0,25	1	0,25	0,25	0,5
Élet	0,25	0,25	1	0,25	0
Egészség	0,25	0,25	0,25	1	0
Nem-élet	0,25	0,5	0	0	1

13. táblázat: A Szolvencia II kockázati moduljai közötti korrelációs mátrix

Változó	Jelentés
C_t	Biztosítási összeg t. időpontban
x	Belépési kor
T	Biztosítás tartama
r	Évesített kockázatmentes hozam (konstans)
t	*t*. időpont ($t = 0, 1, \dots, T$)
q_x	Halálozási valószínűség *x* évesen
p_x	Túlélési valószínűség *x* évesen
$A^{(r)}_{x:T]}$	Egyszeri díjas vegyes biztosítás 1 forintra jutó díja
$\pi(C_t)$	C_t összegű kifizetés díja *t* = 0 -ban
μ_j	$\frac{\eta\gamma_j - i}{1 + i}$
η	A többlethozam növekedését leíró koefficiens
γ_j	$u^{N-j}d^j - 1$
i	Technikai kamat
σ	Volatilitás paraméter a BSM modellben
ρ	Visszavásárlási büntetés (százalékban)

14. táblázat: A 2.2. fejezetben bemutatott képletekben lévő változók jelentése

Távolság	Formula
Euklideszi	$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$
Csebisev	$d(p,q) = \max_i(\lvert q_i - p_i \rvert)$
Koszinusz hasonlóság	$s(x_i, x_k) = \cos\left(\angle(x_i, x_k)\right) = \frac{\sum_{j=1}^{D} x_i^{(j)} x_k^{(j)}}{\sqrt{\sum_{j=1}^{D}\left(x_i^{(j)}\right)^2}\sqrt{\sum_{j=1}^{D}\left(x_k^{(j)}\right)^2}}$
Mahalanobis távolság	$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$; S kovariancia mátrixszal

15. táblázat: A k-NN eljárásban leggyakrabban használt távolságok kiszámítási módjai

Eredeti kategória	**Új kategória**
Elementary Education	Elementary
Primary School	Elementary
None	Elementary
High School	UnderGraduate
Two-year Degree	UnderGraduate
Middle School	UnderGraduate
Graduate	Graduate
Post-Graduate	Graduate
Doctors Degree	Graduate
Doctors Degree and Upper	Graduate

16. táblázat: education változó eredeti és új értékei

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.99392	0.03392	-58.787	< 2e-16	***
ApeCat300-410	0.11635	0.01502	7.747	9.43E-15	***
ApeCat410-518	-2.32377	0.02702	-85.988	< 2e-16	***
ApeCat518-639	-0.99259	0.01816	-54.659	< 2e-16	***
ApeCat639-900	-1.15396	0.01887	-61.162	< 2e-16	***
ApeCat900+	-0.78788	0.01854	-42.499	< 2e-16	***
payment_frequencyannually	0.7829	0.02976	26.308	< 2e-16	***
payment_frequencyquarterly	0.25945	0.07719	3.361	0.000776	***
payment_frequencybiannually	0.43207	0.09512	4.543	5.56E-06	***
payment_methodBank Transfer	0.25187	0.01316	19.146	< 2e-16	***
riskcommencing_year2013	0.76096	0.0211	36.071	< 2e-16	***
riskcommencing_year2014	1.12312	0.02147	52.312	< 2e-16	***
riskcommencing_year2015	1.55486	0.02269	68.528	< 2e-16	***
riskcommencing_year2016	1.55117	0.024	64.639	< 2e-16	***
agency_regionAgencyRegion02	0.36216	0.0323	11.211	< 2e-16	***
agency_regionAgencyRegion03	-0.18033	0.03332	-5.412	6.24E-08	***
agency_regionAgencyRegion04	-0.0206	0.02038	-1.011	0.312076	
agency_regionAgencyRegion05	0.1523	0.0434	3.509	0.00045	***
agency_regionAgencyRegion06	-0.16667	0.02566	-6.496	8.27E-11	***
agency_regionAgencyRegion07	-0.0055	0.02962	-0.186	0.852715	
agency_regionAgencyRegion08	0.13068	0.01618	8.074	6.78E-16	***
agency_regionAgencyRegion09	-0.0646	0.0226	-2.858	0.004258	**
agency_regionAgencyRegion10	0.17415	0.05216	3.339	0.000841	***
agency_regionAgencyRegion11	-0.16186	0.04068	-3.979	6.93E-05	***
distribution_channelJV	0.1669	0.0165	10.112	< 2e-16	***
distribution_channelBank_1	0.11041	0.0152	7.262	3.82E-13	***
distribution_channelDIRECT	-1.18394	0.05135	-23.055	< 2e-16	***
distribution_channelBank_2	-0.16892	0.02859	-5.909	3.45E-09	***
genderM	0.31107	0.01159	26.833	< 2e-16	***
marital_stateSingle	0.30778	0.01228	25.058	< 2e-16	***
educationUnderGraduate	0.06947	0.0158	4.398	1.09E-05	***
educationGraduate	-0.10683	0.01695	-6.304	2.89E-10	***
AgeCat30-45	-0.15193	0.01568	-9.689	< 2e-16	***
AgeCat45-60	-0.37195	0.01802	-20.642	< 2e-16	***
AgeCat60+	-0.14893	0.02567	-5.801	6.60E-09	***

17. táblázat: Az első éves törlések logisztikus regressziós modellének eredményei

	GVIF	Df	GVIF^(1/(2*Df))
ApeCat	1.534026	5	1.043718
payment_frequency	1.281956	3	1.042267
payment_method	1.310373	1	1.144715
riskcommencing_year	1.337498	4	1.037019
agency_region	1.644959	10	1.025198
distribution_channel	1.894765	4	1.083164
gender	1.034097	1	1.016905
marital_state	1.176893	1	1.084847
education	1.126285	2	1.030178
AgeCat	1.329814	3	1.048653

18. táblázat: VIF értékei az első éves törlési modellben

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.5275	0.12587	-4.191	2.78E-05	***
ApeCat490-598	-1.56458	0.04286	-36.504	< 2e-16	***
ApeCat598-656	-2.31646	0.05719	-40.504	< 2e-16	***
ApeCat656+	-1.61924	0.04195	-38.6	< 2e-16	***
payment_frequencyannually	0.9078	0.14958	6.069	1.29E-09	***
payment_frequencyquarterly	0.58795	0.19667	2.99	0.002793	**
payment_frequencybiannually	0.40588	0.43638	0.93	0.352314	
payment_methodBank Transfer	-0.83474	0.03401	-24.545	< 2e-16	***
riskcommencing_year2008	0.3667	0.04624	7.931	2.17E-15	***
riskcommencing_year2009	0.39984	0.04508	8.87	< 2e-16	***
riskcommencing_year2011	0.40824	0.04715	8.658	< 2e-16	***
agency_regionAgencyRegion01	-0.47968	0.12662	-3.788	0.000152	***
agency_regionAgencyRegion02	-0.51381	0.14239	-3.608	0.000308	***
agency_regionAgencyRegion03	-0.84904	0.14484	-5.862	4.57E-09	***
agency_regionAgencyRegion04	-0.72976	0.1279	-5.706	1.16E-08	***
agency_regionAgencyRegion05	-0.9096	0.16702	-5.446	5.15E-08	***
agency_regionAgencyRegion06	-0.66578	0.1323	-5.032	4.84E-07	***
agency_regionAgencyRegion07	-0.54276	0.14007	-3.875	0.000107	***
agency_regionAgencyRegion08	-0.69385	0.12283	-5.649	1.62E-08	***
agency_regionAgencyRegion09	-0.77416	0.12795	-6.051	1.44E-09	***
agency_regionAgencyRegion11	-0.82025	0.15531	-5.281	1.28E-07	***
educationUnderGraduate	0.13066	0.04014	3.255	0.001132	**
educationGraduate	0.32386	0.04393	7.373	1.67E-13	***

19. táblázat: Az ötödik éves törlések logisztikus regressziós modellének eredményei

	GVIF	Df	GVIF^(1/(2*Df))
ApeCat	1.173194	3	1.026979
payment_frequency	1.017976	3	1.002974
payment_method	1.155171	1	1.074789
riskcommencing_year	1.148987	3	1.023417
agency_region	1.158686	10	1.007392
education	1.109807	2	1.026389

20. táblázat: VIF értékei az ötödik éves törlési modellben

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.36674	0.23227	-1.579	0.114356	
ApeCat0-500	-0.07748	0.04994	-1.552	0.12078	
ApeCat500-700	-0.19443	0.04921	-3.951	7.78E-05	***
payment_frequencymonthly	-0.70942	0.19336	-3.669	0.000244	***
payment_frequencyquarterly	-0.36186	0.26857	-1.347	0.177872	
payment_frequencybiannually	-0.83426	0.49286	-1.693	0.090514	.
payment_methodBank Transfer	-0.96859	0.04282	-22.623	< 2e-16	***
riskcommencing_year2004	-0.15607	0.11599	-1.345	0.178469	
riskcommencing_year2005	-0.37815	0.11414	-3.313	0.000923	***
riskcommencing_year2006	-0.52173	0.11565	-4.511	6.44E-06	***
agency_regionAgencyRegion01	-0.47209	0.07998	-5.903	3.58E-09	***
agency_regionAgencyRegion02	-0.48873	0.12261	-3.986	6.72E-05	***
agency_regionAgencyRegion03	-0.35448	0.08907	-3.98	6.90E-05	***
agency_regionAgencyRegion05	-0.37339	0.13186	-2.832	0.00463	**
agency_regionAgencyRegion06	-0.36153	0.08771	-4.122	3.76E-05	***
agency_regionAgencyRegion07	-0.38375	0.09863	-3.891	9.99E-05	***
agency_regionAgencyRegion08	-0.31877	0.05576	-5.717	1.09E-08	***
agency_regionAgencyRegion09	-0.36914	0.06726	-5.488	4.06E-08	***
agency_regionAgencyRegion10	-0.56421	0.17185	-3.283	0.001027	**
agency_regionAgencyRegion11	-0.94574	0.12062	-7.841	4.48E-15	***
genderM	-0.27127	0.04792	-5.661	1.50E-08	***
marital_stateSingle	-0.20332	0.05345	-3.804	0.000142	***
educationUnderGraduate	0.23609	0.04893	4.825	1.40E-06	***
educationGraduate	0.33871	0.05255	6.445	1.16E-10	***
occupationBlue_collar_employee	0.19481	0.06048	3.221	0.001277	**
occupationPink_collar_employee	0.37807	0.06715	5.63	1.80E-08	***
occupationOther	0.04159	0.05117	0.813	0.4164	
AgeCat45-60	0.29583	0.04384	6.748	1.49E-11	***
AgeCat60+	1.51945	0.05602	27.123	< 2e-16	***

21. táblázat A tizedik éves törlések logisztikus regressziós modellének eredményei

	GVIF	Df	GVIF^(1/(2*Df))
ApeCat	1.213918	2	1.049657
payment_frequency	1.02656	3	1.004378
payment_method	1.255285	1	1.120395
riskcommencing_year	1.1165	3	1.018536
agency_region	1.195113	10	1.008952
gender	1.518627	1	1.232326
marital_state	1.044837	1	1.022173
education	1.237793	2	1.05478
occupation	1.813889	3	1.104337
AgeCat	1.124222	3	1.019707

22. táblázat: VIF értékei a tizedik éves törlési modellben