

OTDK-dolgozat

2023

BUDAPESTI CORVINUS EGYETEM
MARKETING- ÉS KOMMUNIKÁCIÓTUDOMÁNYI INTÉZET



**KERESZT- ÉS FELÜLÉRTÉKESÍTÉSI ELŐREJELZÉS A MAGYAR
TELEKOMMUNIKÁCIÓS SZÉKTORBAN –
VÉLETLEN ERDŐ MÓDSZERREL**

**CROSS- AND UPSELLING PREDICTION IN THE HUNGARIAN
TELECOMMUNICATION INDUSTRY -
USING RANDOM FOREST ALGORITHM**

Tóth Zoltán

Konzulens: Dr. Kenesei Zsófia

Kézirat lezárásának dátuma: 2021. 04. 12.

Tartalomjegyzék

I.	Bevezetés	1
II.	Szakirodalom elemzés.....	5
II.1.	Szakirodalmi elemzés keretrendszere	5
II.2.	Lojalitás.....	5
II.2.1.	A lojalitás jelentősége	7
II.2.2.	Ágazati mérések	8
II.3.	Kereszt- és felülértékesítés	10
II.3.1.	Definíció és általános jellemzők	10
II.3.2.	Telekommunikációs és pénzügyi szektorban.....	14
II.4.	Fogyasztómegtartás és lemorzsolódás	16
II.4.1.	Meghatározása és jelentősége	16
II.4.2.	A telekommunikációs szektorban	18
II.4.3.	A pénzügyi szektorban.....	19
II.5.	Adatbázis-marketing	21
II.5.1.	Adatbázis marketing.....	21
II.5.2.	Adatbányászat	22
II.5.3.	Mesteréges intelligencia.....	23
II.5.4.	A telekommunikációs és pénzügyi szektorban	24
II.6.	A kereszt- és felülértékesítés előrejelzésére alkalmazott prediktív módszerek bemutatása.....	26
II.6.1.	NPTB	26
II.6.2.	Véletlen erdő (Random Forest).....	29
II.6.3.	Logisztikus regresszió	31
III.	A kutatás kérdései és módszere.....	31
III.1.	Kutatási célok.....	31
III.2.	A vizsgált adatbázis	33
III.3.	Az elemzéshez használt program és módszer	34

IV.	Kvantitatív kutatás:	36
IV.1.	KNIME folyamat ábra:	36
IV.2.	Adatok importálása és előkészítése	37
IV.3.	Kiszűrt jellemzők:	37
IV.4.	Hiányzó értékek kezelése.....	38
IV.5.	Kiugró értékek kezelése	40
IV.6.	Tanuló és teszt adatok szétválasztása	43
IV.7.	Alulmintázás	44
IV.8.	Véletlen erdő	45
V.	Kutatási eredmények	48
V.1.	A modell Pontosságának (Precision) és Felidézésének (Recall) kiértékelése 48	
V.2.	Legjelentősebb vágó kritériumok	50
V.3.	Legkevesbé jelentősebb vágó kritériumok.....	54
V.4.	Eredmények kiértékelése	55
V.5.	Megjegyzések	56
VI.	Magyar Telekom esettanulmány	57
VII.	További előnyök:	63
VIII.	Legnagyobb kihívások és a modell korlátai	65
VIII.1.	Legnagyobb kihívások	65
VIII.2.	A modell korlátai és fejlesztési lehetőségek	67
VIII.3.	A Pontosság és a Felidézés érzékenységének összehasonlítása	70
IX.	Összefoglalás:.....	71
X.	Irodalomjegyzék.....	75
XI.	Mellékletek:.....	82
XI.1.	A modell beállításai:	82

Táblázatjegyzék

1. Táblázat A hiányzó értékek kezelésére használt módszerek (numerikus értékek esetén) összehasonlítása.	39
2. Táblázat A hiányzó értékek kezelésére használt módszerek (szöveges értékek esetén) összehasonlítása.	40
3. Táblázat A kiugró értékek kezelésére során alkalmazott R értékek eredményeinek összehasonlítása.	41
4. Táblázat A kiugró értékek kezelése során alkalmazott k értékek eredményeinek összehasonlítása.	41
5. Táblázat A kiugró értékek cseréje során alkalmazandó módszer kiválasztását segítő táblázat.	42
6. Táblázat A kiugró értékek során a cserélni kívánt értékek kiválasztását segítő táblázat.	42
7. Táblázat A megfelelő tanító-teszt arány kiválasztását segítő táblázat.	43
8. Táblázat A megfelelő mintavételi mód kiválasztását segítő táblázat.	44
9. Táblázat A megfelelő alulmintázási arány kiválasztását segítő táblázat.	45
10. Táblázat A megfelelő mérési módszer kiválasztását segítő táblázat.	46
11. Táblázat A csomópontok minimális számának kiválasztását segítő táblázat.	46
12. Táblázat A döntési fák maximális mélységének meghatározását segítő táblázat. .	47
13. Táblázat A véletlen erdő méretének meghatározását segítő táblázat.	47
14. Táblázat A végső modell lefutási eredményeit összefoglaló táblázat.	48
15. Táblázat A véletlen erdő által legfontosabbnak talált jellemzők.	51
16. Táblázat A különböző forgatókönyveket bemutató táblázat.	58
17. Táblázat A projekt költségeit (a jutalék költségeken kívül) összesítő táblázat. Forrás: Saját szerkesztés.	60
18. Táblázat Az egy értékesítési migrációval képződő bevétel meghatározását segítő táblázat.	61
19. Táblázat A vállalat hívás számait összefoglaló táblázat.	61
20. Táblázat A vállalat pénzügyi eredményeit összefoglaló táblázat.	61
21. Táblázat A Felidézés változásának hatása a profitra.	70
22. Táblázat A Pontosság változásának hatása a profitra.	71

I. Bevezetés

„A fogyasztók nem tudják mit akarnak, amíg meg nem mutatjuk nekik.”

Steve Jobs¹

Steve Jobs az 1980-as években elhangzott ismert mondatát azért választottam dolgozatom nyitásaként, mivel már közel 40 évvel ezelőtt felhívta figyelmet a marketingkutatókban és a termékfejlesztésben rejlő egyik problémára, miszerint túlságosan arra támaszkodunk, hogy mit mondanak nekünk a fogyasztóink, hogy mit szeretnének, annak ellenére, hogy sokszor ezzel ők maguk sincsenek tisztában. Ez a jelenség pedig ötvöződik azzal, hogy a fogyasztók mellett sokszor a saját rezervációs árakkal sincsenek tisztában, mindez tehát újfajta megoldásokat követel. (Danyi et al, 2020)

A heurisztikán alapuló értékesítés egyre nagyobb problémákba ütközik két oldalról is. Egyrészt a fogyasztóknak nem vonzó, ha számukra teljesen érdektelen termékekkel kapcsolatban keresik meg őket, amely azt az érzést keltheti bennük, hogy a vállalat nem foglalkozik velük igazán, csak bármit vagy adott esetben mindent el akar nekik adni. Emellett a vállalat szempontjából is problémás ez, hiszen a telefonos értékesítés alapvetően drága és amennyiben hozzá tesszük a hívási arányok sikerességét, azt is láthatjuk, hogy jelen állapotában egyáltalán nem hatásosan használják ki a vállalatok erőforrásaikat. Már a koronavírus járvány előtt is a marketing költség csökkentése volt megfigyelhető nagyvállalatok esetében, jelen helyzetben ez a tendencia tovább erősödött. Míg a Magyar Telekom Nyrt. 2015-ben több, mint 13 milliárd² forintot költött marketingre, addig 2019-ben már csak ~10,6 milliárdot³, történt mindez a bevételek ~656 milliárd forintról ~666 milliárd forintra történő növekedése mellett.

Mindezen felül az Amerikai Egyesült Államokban már törvényi védelem is alakult arra vonatkozóan, hogy a „zaklató” telemarketingtől megvédjék a fogyasztókat, amely szintén az értékesítésben történő változtatásokat sürgeti.⁴

Dolgozatom témájának mindezek alapján a kereszt- illetve felülértékesítés területét választottam, mivel úgy ítélt meg, hogy ennek a területnek az utóbbi években hányatott sorsa volt. A lojalitás növelése, a panaszkezelési folyamatok javítása vagy éppen a lemorzsolódás

¹ <https://hbr.org/2012/04/the-real-leadership-lessons-of-steve-jobs>

² https://www.telekom.hu/static-tr/sw/file/IFRS_Csoport_2016.pdf

³ https://www.telekom.hu/static-tr/sw/file/IFRS_Csoport_20191231_AGM.pdf

⁴ <https://www.cmswire.com/digital-marketing/meet-the-new-telemarketing-proactive-positive-and-personalized/>

csökkentése, a digitális marketing előretörése esetén kevés szó esett erről a területről, pedig előbbiek is kapcsolódnak a kereszt-és felülértékesítéshez. Történt mindez annak ellenére, hogy Jeff Bezos napjaink leggazdagabb embere, a világ egyik legnagyobb vállalatának, az Amazonnak az alapítója 2006-ban nyilatkozott arról, hogy a vállalat értékesítésének több, mint az egyharmada származott keresztértékesítésből, amely jól mutatja a területben rejlő lehetőségeket.⁵

Dolgozatomban kiemelten a telemarketinggel összeköthető kereszt- és felülértékesítéssel fogok foglalkozni, mivel úgy érzem ezen a területen van a legnagyobb szükség a jelenlegi megoldások át- és tovább gondolásra. A telemarketing jelentősége még napjainkban is kiemelkedő, az ügyfélélmény fontos részét teheti ki az ügyfélszolgálattal való kapcsolat, és amennyiben jól hasznosítjuk, vállalatunk egyik legnagyobb értékévé válhat az ügyfélszolgálat.⁶

Ezen terület támogatásához kapcsolódóan vizsgálom a gépi tanuló algoritmusok világát, azon belül is a véletlen erdő modellben rejlő lehetőségeket. A gépi tanuló algoritmusok használata a XXI. század legnagyobb trendjei közé tartozik, az adattudományt a „század legszexibb munkakörének” nevezi a Harvard Business Review⁷. A gépi tanulást rengeteg szektorban hasznosítják sikeresen: gondolhatunk különböző képfelismerési rendszerekre, az online hirdetési piacra vagy éppen arra, hogy milyen zenét ajánl nekünk a Spotify⁸. A gépi tanuló algoritmusok mindennapjaink részévé váltak anélkül, hogy tudomást szereztünk volna róluk.

Dolgozatom során kiemelten a hazai telekommunikációs piacot vizsgálom, melyet több szempont miatt választottam. Egyrészt egy hazai telekommunikációs vállalatnál dolgozom, emiatt belülről érzékelttem a probléma aktualitását és fontosságát, továbbá ezáltal lehetőségem volt valós adatbázisból származó adatokon dolgozni. Másik oldalról pedig a telekommunikációs szektor egy érdekes választás annak tekintetében, hogy alapvetően a szolgáltatók 4-5 (mobil, TV, internet, vezetékes telefon, készülék értékesítés) szolgáltatással rendelkeznek, melyeket különböző csomagokba rendelve kínálnak, ezáltal kiváló lehetőséget jelentenek a kereszt- és felülértékesítés számára. Továbbá olyan szektorról beszélhetünk, ahol alapvetően hazai és nemzetközi szinten is nagyvállalatok vannak a piacon többek között a

⁵ <https://predictableprofits.com/amazon-can-teach-cross-selling/>

⁶ <https://customerthink.com/why-the-call-center-matters-more-than-ever/>

⁷ <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

⁸ <https://elitedatascience.com/machine-learning-impact>

magas belépési korlátok miatt, ezen nagyvállalati háttérnek köszönhetően pedig a gépi tanuló algoritmusok kiépítésére jó lehetőség nyílik, mivel rengeteg adat keletkezik, amelyeknek nagy része valamilyen formában valószínűleg jelenleg is gyűjtésre kerül.

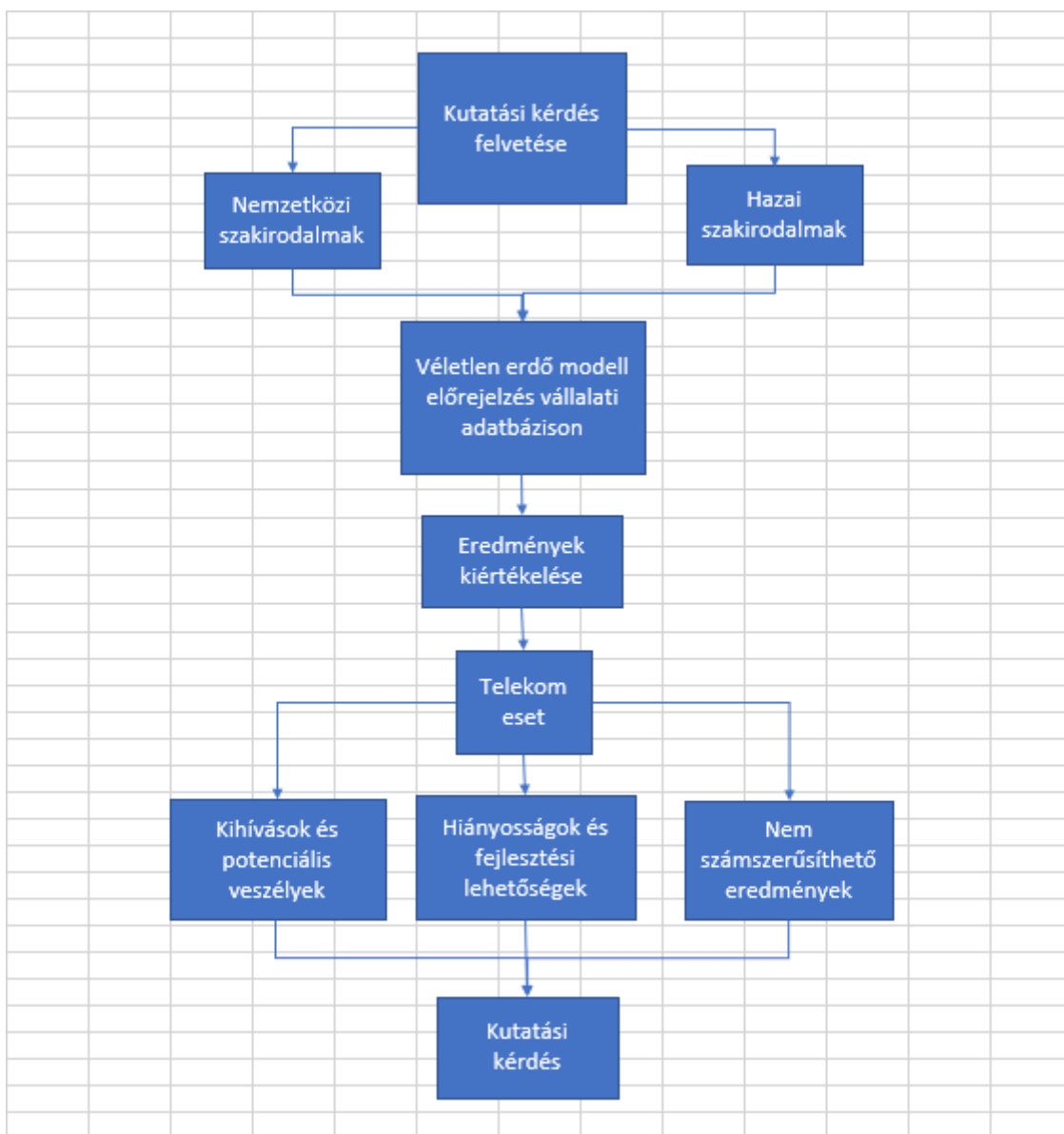
Személyes motivációm továbbá, hogy gyermekkorom óta érdeklődést mutatok a számok és a statisztika világa iránt. Mikor megkaptam az első nagy atlaszomat másokkal ellentétben nem csak a szép képeket és térképeket nézegettem, hanem sokkal inkább az érdekelt, hogy hányan laknak egy országban vagy éppen mekkora az adott ország területe. Ezen felbuzdulva füzetben kezdtem írni a legnépesebb országokat, amely listában rendszeresen kerültek áthúzásra országok és változott a sorrend, ahogy egyre jobban haladtam előre az atlaszban. Bár nem nevezném effektívnek a módszerem, azonban már gyermekkoromban adatbázist építettem és választ adtam egy sorba rendezési problémára.

Mindezek alapján választottam dolgozatom kutatási kérdésévé a következőt: *„A véletlen erdő módszer segítségével milyen módon és mekkora mértékben segíthető elő a hazai telekommunikációs szektorban a felül- és keresztértékesítés?”*

Dolgozatom első részét egy részletes hazai és nemzetközi szakirodalmakat felölelő irodalomelemzéssel fogom kezdeni, ahol a marketing oldaláról járom körbe a releváns témákat, ezt követően pedig válogatott előrejelző módszerek kerülnek bemutatásra. Ezután részletezem a vállalati adatbázishoz kapcsolódó modellépítésemet és értelmezem ennek eredményeit. A dolgozatot úgy építettem fel, hogy az útmutatóként is szolgálhat más vállalatok számára.

Ezt követően kitérek a modell korlátaira, fejlesztési lehetőségeket és különböző nem számszerűsíthető eredményeket szemléltetek, majd dolgozatom a legfontosabb gondolatok összegzésével zárom.

Dolgozatom gerincét az alábbi keretrendszer adja:



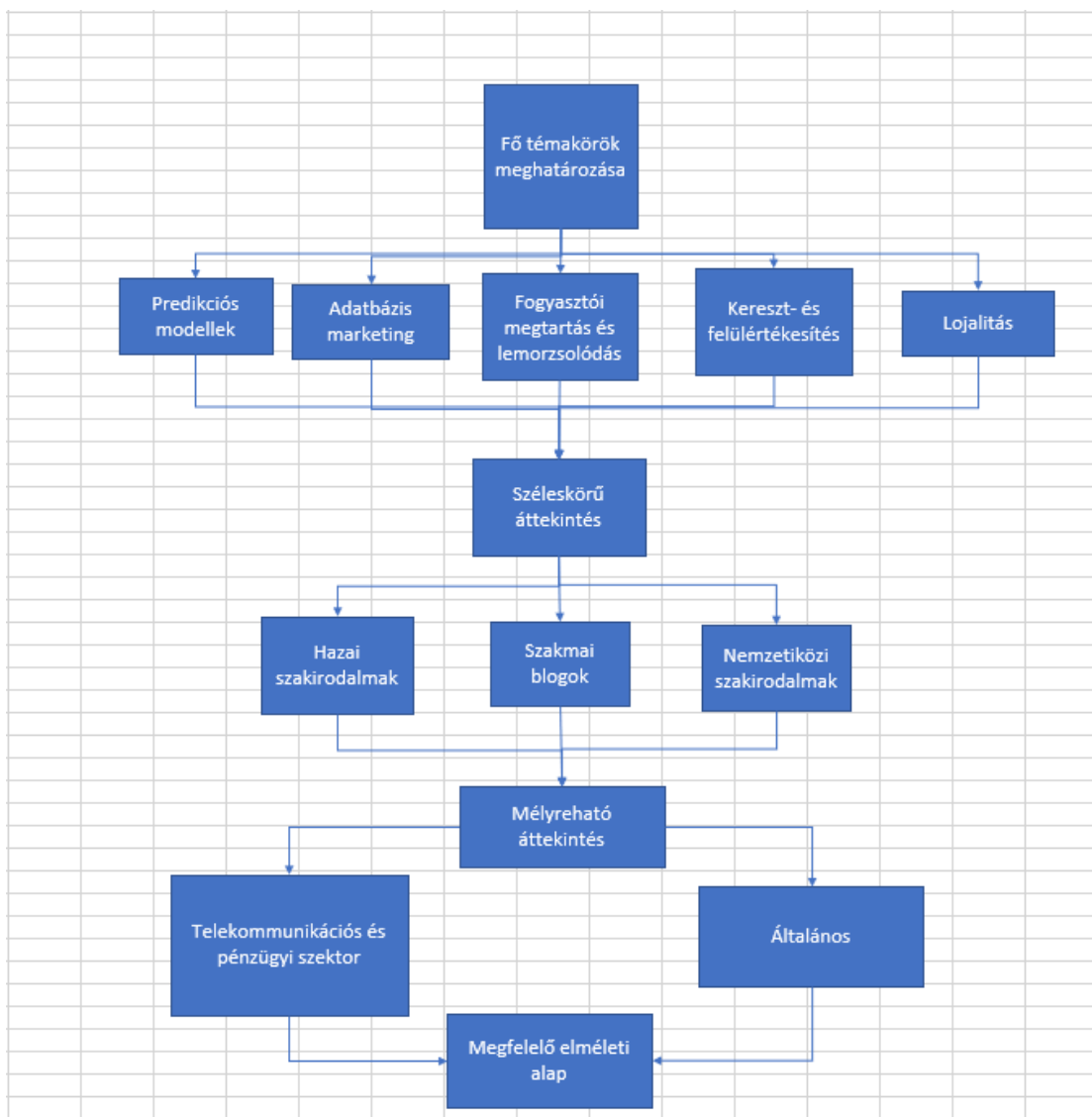
1. ábra Kutatásom keretrendszere.

Forrás: Saját szerkesztés.

II. Szakirodalom elemzés

II.1. Szakirodalmi elemzés keretrendszere

A következő szakirodalmi elemzés az alábbi keretrendszer alapján készült:



2. ábra Szakirodalmi áttekintésem keretrendszere.

Forrás: Saját szerkesztés.

II.2. Lojalítás

A fogyasztói lojalítás, illetve a fogyasztói lojalitásprogramok évtizedek óta az üzleti élet és a marketing fontos részét képezik, nincs ez másként napjainkban sem, mikor is a fogyasztókért való harc kiélezettebb, mint valaha. Mindenekelőtt definiálni kell a fogyasztói

lojalitást, meghatározni, mit is értünk alatta, és mivel több definíció is létezik, ezért meghatározom azt, amely ezt a tudományos munkát végig kíséri.

Horváth és Bauer (2016) szerint a lojalitás egyenlő az újravásárlással, habár úgy gondolom, hogy az újravásárlás egy jó és fontos mérőeszköz a lojalitásnál, ettől függetlenül lehet egy fogyasztó lojális újravásárlás nélkül is, illetve „nem lojális” annak ellenére, hogy újravásárolja az adott terméket. Rekettye, Töröcsik és Hetesi (2016) már kétfelé bontja a (B2B) lojalitás kérdését: meghatározzák a magatartási lojalitást, azaz azt, mikor megtörténik az újravásárlás ugyanattól partnertől, (ez gyakorlatilag megegyezik a Horváth-Bauer féle definícióval) és az attitűdinális lojalitást, amely ennek már egy magasabb szintű változata, mikor is az újravásárlás mellett / alatt érzelmi kötődés is kialakul a felek között.

Oliver definícióját használja Keller és Kotler (2016), akik a lojalitást szintén nem teszik egyenlővé az egyszerű újravásárlással, hiszen kikötik, hogy lojalitás esetén az újravásárlási hajlandóság mélyen gyökerező, és ezáltal a lojalitás képes áthatolni az erőteljes külső hatásokon (azaz más versenytárs vállalatok marketing tevékenységein). Ezen három definíció mellett egy utolsó negyediket említek meg, amely után összegzem ezeket: Agárdi (2017) szerint a lojalitásról akkor beszélhetünk, ha az adott fogyasztó egy vállalatot konzisztensen előnyben részesít másokkal szemben, és ez a lojalitás az újravásárlás mellett megjelenhet abban is, hogy más vállalatok tevékenysége nem ösztönzi őt a náluk való vásárlásra, azaz itt is megjelenik a külső hatásokon való elhatárolódás.

Kenesei és Kolos (2014, 85. old.) meghatározása alapján: *„Tágabb értelemben a lojalitás a vevő ragaszkodása, kötődése egy vállalat termékeihez, szolgáltatásaihoz vagy akár alkalmazottaihoz. Szűkebb értelemben a lojalitás az a fogalom, amely összefoglalja a szolgáltatóval szembeni attitűdöt és bizonyos magatartásformákat.”* Kemény (2015, 80. old.) pedig a fellelhető lojalitás definíciókat összegezte a következőképpen: *„Összefoglalásként megállapítható, hogy a lojalitást sok esetben egy márka/termék többszöri újravásárlásaként, vagy boltok esetében egy üzlet többszöri újravásárlásaként definiálják”.*

A felsorolt meghatározások alapján én a dolgozatomban a következőképpen definiálom a lojalitást: Lojális fogyasztónak nevezzük azt a fogyasztót, aki az adott vállalattal kapcsolatosan hosszú távon támogatólag lép fel, és teszi ezt olyan módon, hogy a vállalatot ezzel üzleti haszonhoz segíti.

II.2.1. A lojalitás jelentősége

A lojalitás definiálása után fontos meghatározni, hogy a lojalitás miért fontos része a XXI. századi marketingnek (is). Agárdi (2017) a fogyasztói lojalitást az egyik legértékesebb versenyelőnynek tekinti. Kenesei és Cserdi (2018) ennél már valamelyest árnyaltabban fogalmaznak: megemlítik, hogy mivel nem minden vásárló (és vásárlás) profitábilis, illetve nem ugyanakkora mértékben, ezért fontos meghatározni, hogy melyek azok a vásárlók, akiknek a lojalitása növelendő, és melyek azok a vásárlók, akiket adott esetben egyáltalán nem érdemes megtartania a vállalatnak, mert csak veszteséget termelnek (például egy biztosítótársaság esetén).

A lojalitás kialakulását és növelését több eszköz segítségével is elérhetjük: Heidrich (2017) az alábbiakat emelte ki: tudás (a vállalaton belül az évek által felhalmozott közös tudás, amely más területeken is versenyelőnyt jelenthet), személyre szabott érték létrehozása (azaz ne homogén csoportként tekintsünk az ügyfelekre, igyekezzünk meghatározni olyan profilokat, amelyek se nem túl szűkek, se nem túl tágak, és ezeket egyedi módokon targetáljuk). Elengedhetetlennek tartja az ügyfél szükségletének feltárását, az állandó párbeszédet az ügyfelekkel (fontos tisztában lenni azzal, hogy az ügyfelek mit gondolnak a vállalatról, hogy adott esetben egy-egy promóció vagy történés milyen impulzusokat vált ki a fogyasztóknál és ezeket fontos hosszú távon tárolni és ez alapján kapcsolatot létesíteni velük), és a pozitív egységes élmény nyújtását Heidrich (2017). A felsorolt tényezők között nem szerepel az ár, így bármennyire is gondolnánk triviálisnak, hogy az ár a lojalitás kialakuláshoz vezető egyik legfontosabb tényező, ez nem így van, ezt Danyi et al (2020) kutatása is megállapította.

A Heidrich által említett öt tényezőből azonban az rajzolódik ki, hogy ahhoz, hogy ezeket hosszú távon jól ki tudja használni egy vállalat, szükség van egy jól kialakított és kihasznált egységes vállalati adatbázisra, melyre a későbbiekben rá fogok térni.

A fogyasztói lojalitás és a CRM rendszerek kapcsolatával Piskóti (2016) is foglalkozott egy piackutató-elemző cégek által elvégzett felmérés során. A felmérésben összesen 300 európai és észak-amerikai vállalatot kérdeztek meg arról, milyen szempontok voltak a legfontosabbak egy CRM rendszer bevezetésekor és a leggyakoribb esetek között voltak az általam is említettek: az ügyfelek hűségének növelése, személyre szabott szolgáltatások nyújtása, az ügyfelek jobb ismerete és a legjövedelmezőbb ügyfelek megállapítása.

Rávilágítottam korábban, hogy nem minden vásárló esetén célszerű elérni a lojalitást, azonban összességében a lojalitás növelése szinte kivétel nélkül egy eredményességet növelő

folyamat. Reichheld és Sasser (1990) kutatások során megállapították, hogy a lojalitás 5%-os növelésével egy csomagküldő vállalat esetén a profit 20%-kal nő, egy biztosító társaság esetén 50%-kal, egy hitelkártyákat biztosító társaság esetén pedig 125%-kal. Ebből jól látszik, hogy habár a lojalitás növelése költséges lehet (főleg a kezdeti beruházások esetén), hosszabb távon relatív (százalékban) alacsony lojalitásnövekedés is jelentős, adott esetben óriási profitnövekedéssel járhat, attól függően, hogy milyen szektorokban érdekelt a vállalat.

Az eddigiekből már jól látszódik, hogy a fogyasztói lojalitás milyen fontos részét képezi a vállalati marketingnek, azonban ezt az állítást még két továbbival kívánom erősíteni: egyrészt Keller márkarezonancia piramisával, ahol Keller a márkaépítéssel kapcsolatosan egy piramist határozott meg, melynek legfelső szintjén a rezonancia áll, amely gyakorlatban megfeleltethető a fogyasztói lojalitásnak (Bauer- Kolos, 2017). Ezen kívül a márkázás területén egy másik ismert piramis, a Millward Brown által létrehozott Brand Piramis, melynek a legfelső szintjén éppúgy a lojalitás foglal helyet, ez esetben „kötődés” néven szerepeltetve. (Malhotra-Simon, 2017)

Összességében tehát elmondhatjuk, hogy egy vállalat esetében a profitabilitás növelésére az egyik legjobb módszer a fogyasztói lojalitás növelése. A bevételek növekedése mellett fontos megemlíteni, hogy a magas vásárlói lojalitás a költségeket is csökkenti, hiszen lassabban cserélődnek a fogyasztók, illetve a szájreklám alkalmazása is sokkal jellemzőbb a lojális fogyasztók körében, ezek miatt pedig a marketing költségek lényegesen csökkenthetőek. (Gould, 1995) Ezen előnyök miatt a XXI. században több milliárd dollárt költenek a különböző vállalatok a saját lojalitásprogramjaikra, mindemellett a kutatási szektorban is több tízezer könyv és több százezer szakcikk mutatja a téma fontosságát, (Keiningham et al, 2011) de a legszemléletesebben az mutatja be a lojalitásprogramok monumentalitását, hogy a törzsutas mérföldek tekinthetőek a világ második legnagyobb fizetőeszközének az USA dollár után. (Evanschitzky et al, 2012)

II.2.2. Ágazati mérések

A következőkben a fogyasztói lojalitást, lojalitásprogramokat az infó-telekommunikációs, banki és biztosítási szektorban fogom röviden bemutatni. Ahogy korábban kitértem rá, a különböző szektoroknak más-más sajátosságai vannak a fogyasztói lojalitás területén is és ez ezen szektorok esetén sincs másképp. A biztosítási szektoron belül például egy kutatás során kiderült, hogy a piacra egyre jobban bekerülő fiatalabb generációk sokkal könnyebben váltanak egy-egy jó ajánlat esetén, mint az idősebbek, azaz felértékelődik a

különböző lojalitásprogramok értelme, illetve a személyre szóló ajánlatok létesítése. (Szőke, 2015)

Liu (2008) kutatása során - a kínai telekommunikációs szektor vizsgálatakor megállapította, hogy alapvetően a fogyasztói elégedettség az egyik legfontosabb változó, ami a vásárlói lojalitáshoz vezethet, azonban azt is kimutatta, hogy mivel ez a megállapítás kizárólag múltbéli adatokon alapul, így inkább arra hasznos, hogy megítéljük, a korábbi lépéseink mennyire voltak pozitívak, vagy sem, előrejelzésekre azonban kevésbé alkalmasak. Liu emellett a vállalati imázst fedezte fel, mint az egyik legfontosabb befolyásoló tényező a fogyasztói lojalitással kapcsolatban. Ezek olyan információk, amelyek a magyar piaccal kapcsolatban is igazak lehetnek, hiszen a vevői elégedettség és a vállalati imázs gyakorlatilag szektortól és országtól függetlenül megjelenik a világon, mint az üzlet és a marketing fontos elemei.

Az indiai mobil telekommunikációs szektort a fogyasztói lojalitás szemszögéből vizsgálta Kaur és Soch (2012). Az indiai IKT iparágával kapcsolatban fontos megjegyezni, hogy több szempontból is egyedülálló a világon. Egyrészt a Föld legtöbb telefon előfizetővel rendelkező országa India, másrészt pedig az egyik legolcsóbb szolgáltatással rendelkező, ahol a feltöltőkártyás szolgáltatások vannak nagy túlsúlyban, illetve rengeteg a piaci szereplő (9-12 szolgáltató) és ráadásul az alacsony belépési küszöb miatt ez a szám tovább nőhet. Ez a magyar IKT szektortól nagyban különbözik, hiszen hazánkban gyakorlatilag 3 nagy szereplő szerepel évek óta akár a vezetékes akár a mobilos telekommunikáció esetén és több kisebb szereplő mellettük (általában) regionálisan. Mindezekről függetlenül az indiai telekommunikációs szektorban a kínaihoz és más egyéb szektorokhoz hasonlóan hosszú távon a vásárlói lojalitásban látják az egyik legnagyobb jövőt, mivel így lehetséges csak a versenyben maradás.

Ezen felül Kaur és Soch (2012) egy egyszerű modellt is felvázoltak, amely szerint a vásárlói lojalitáshoz alapvetően két tényező vezet: a vállalati imázs (ahogy Liu is említette), illetve a váltási költségek (amely attól függetlenül, hogy más az indiai és a magyar szektor, szintén jelen van, mint egy fontos tényező hazánkban is, hiszen ezen a területen a szolgáltató vásárlás sokkal időigényesebb tud lenni, mint mondjuk egy élelmiszer márkák közötti váltás esetén). Azonban a kutatásuk itt még nem állt meg, mivel megállapították, hogy ezen két tényezőt négy másik hatás is befolyásolta: a vásárlói elégedettség (ahogy már korábban is megjelent és itt is megjelenik, de szintén nem direkt kapcsolatként), a bizalom és a vállalat iránti elköteleződés (amit a szerző ketté bont affektív és kalkulatív elköteleződésre).

A harmadik és egyben utolsó nemzetközi példa egy tajvani kutatás (Chen – Cheng, 2012), amely során egy előre felállított modellel kapcsolatosan vizsgáltak nyolc darab hipotézist, melynek a végén megállapították, hogy a fogyasztói lojalításra a vásárlói elégedettség és az érzékelt érték (ez indirekt módon is hat a fogyasztói lojalításra) vannak közvetlen hatással, és ez a modell is jól mutatja, hogy jelenleg nem lehet konkrét modellt meghatározni, amely minden piacra igaz lenne. A vizsgálatot mindig az adott helyzethez kell igazítani és kellő fontossággal szükséges felmérni, hogy milyen változók, milyen arányban lehetnek érdekeltek.

Magyarországi példaként Agárdi és Gyulavári (2017) végeztek kutatást, melynek során a magyarországi banki, illetve biztosítási szektorban vizsgálták a CRM-stratégiákat és ez alatt a fogyasztói lojalitást is. Megemlítik az előbbi nemzetközi kutatásokkal egyetemben, hogy a CRM fejlesztések a pénzügyi szektorban (és ehhez kötődően a fogyasztói lojalitás) nagy hatással vannak a vállalatok versenyképességére. A szerzők itt egy olyan megállapításra jutottak, hogy alapvetően egy körkörös folyamatról van szó, amelyben a fogyasztói lojalitásnak is szerepe van. Ezen modell alapján a fogyasztói lojalításra hatással vannak a személyre szabott ajánlatok és a magasabb szolgáltatási színvonal. A kutatásuk során magyarországi CRM-vezetők körében végeztek felmérést, ezen felmérés alatt kiderült, hogy a lojalitás növelése kiemelt szerepet tölt be a vezetők, és kiemelten a biztosító társaságok vezetői körében. Emellett megállapították, hogy a vezetők a lojalitás esetén az érzelmi kötődés kialakítását (azaz az attitűdinális lojalitást) kifejezetten nagy kihívásnak jelezték, mivel úgy ítélték meg, hogy ez egy nem megvásárolható összetevő (ez összecseng azzal a korábbi megállapítással, hogy az alacsony árak nem feltétlen jelentenek növekedő hűséget). Emellett azonban a CRM-vezetők esetében is előjött, hogy a fogyasztói elégedettség növeli a lojalitást, ami jól mutatja, hogy ez a változó sem ország specifikus, hanem (minimum) a szektoron belül általánosan igaz. Érdekes megállapítás volt még a kutatásban emellett, hogy a vezetők jelezték, hogy a különböző csatornákról származó kvalitatív információk- amelyek például egy hibabejelentés, vagy egy call centeres ajánlattétel során történnek- sok olyan információt hordozhatnak magukban, amelyek a tranzakciós adatokkal összevetve értékes tudást képezhetnek.

II.3. Kereszt- és felülértékesítés

II.3.1. Definíció és általános jellemzők

Szerencsénkre a keresztértékesítés és a felülértékesítés esetén sokkal egyszerűbb és kevésbé összetett definícióval kerülünk szembe, mint a fogyasztói hűség esetén. Piskóti (2016) szerint a keresztértékesítés esetén arról beszélünk, hogy a bevétel növelése érdekében

kiegészítő termékeket vagy szolgáltatást kínálunk a vásárlóknak, a felülértékesítés esetén pedig egy drágább (magasabb) szintű szolgáltatást értékesítünk.

Elengedhetetlen, hogy vállalaton belül az egyes dolgozók és részlegek nyitottak legyenek egymás felé, illetve kellő ismerettel rendelkezzenek a vállalat profiljával kapcsolatosan. Ennek hiányában könnyen előfordulhat, hogy a keresztértékesítés a belső ellentétek miatt megbukik (például X részleg jobban akar teljesíteni Y-nál), vagy éppen azért nem jön létre, mert egy adott részleg munkatársai nincsenek tisztában azzal, hogy egy másik részleg egyáltalán milyen termékeket, milyen konstrukcióban értékesít. (Poór, 2017)

Heidrich (2017) tanulmányából kiderül, hogy a már lojalitás témakörében is megemlített profilozás itt is fontos funkcióval rendelkezik, hiszen elengedhetetlen, hogy naprakész információnk legyen arról, hogy kik az aktív, folyamatosan vásárló ügyfeleink, mivel őket érdemes keresztértékesítéssel vagy felülértékesítéssel megkeresni. Heidrich ezt a felvetését a lojalitással való összekötéssel indokolja: úgy gondolja, hogy azon vevők, akik folyamatosan vásárolnak, már rendelkeznek valamilyen fajta kötődéssel a vállalat felé, így van értelme nekik új termékeket vagy szolgáltatást is ajánlani.

Megítélésem alapján Heidrich és Poór állításai is helytállóak, és mindkettőben közös pont, hogy kiemelik azt, hogy az információ kellően pontosan legyen rögzítve és legyen elérhető az adott felek számára, beszéljünk itt akár a vállalat saját tulajdonságairól, termékpalettájáról vagy az ügyfelek profiljairól.

Nemzetközi szinten sokan vizsgálták a keresztértékesítést: Knott et al (2002) hozta létre az egyik legátfogóbb kutatást a keresztértékesítéssel és az NPTB modellel (amellyel később részletesen fogok foglalkozni) kapcsolatosan. Knott megállapította, hogy keresztértékesítés során két nagy kérdésre kell alapvetően választ adni: az egyik az, hogy kik azok az ügyfelek, akiket célozni érdemes, illetve melyek azok a termékek a palettánkról, amelyekkel megérhetné őket célozni. Ezen kérdések azért is ilyen kritikusak, mert nem célozhatunk minden fogyasztót minden termékkel több okból sem: egyrészt attól, hogy egy fogyasztót érdekelt egy vagy akár több termékünk és meg is vásárolta őket, attól még nem biztos, hogy a többi termékünk is fogja (például egy fiatal 18 évesnek hiába van hitelkártyája és diákhitel akár, attól még életbiztosítást nem biztos, hogy akar kötni), és amennyiben olyan termékajánlatokkal bombázzuk, amelyek neki irrelevánsak, akkor akár az is előfordulhat, hogy teljesen elpártol tőlünk. Emellett maga a folyamat, amíg minden fogyasztót elérnénk minden termékkel kapcsolatosan nagyon

erőforrásigényes is lenne (mind pénzben és időben), és a végén minden valószínűség szerint egy rendkívül alacsony sikerességi rátánk jönne létre.

Egy sikeres keresztértékesítési kampány esetén fontos, hogy tisztában legyünk azzal, hogy a vásárlóink mely termékeket preferálják és ehhez célszerű felhasználni minden rendelkezésre álló adatunkat, csakúgy, mint a fogyasztói hűség növelésének esetén. (Akcura és Srinisavan, 2005)

Azonban felhívták a figyelmet arra is, hogy ez a folyamat magával hordoz egy kritikus problémát is, amely 2005 óta még sokkal aktuálisabb és fontosabb lett: az adatbiztonság kérdését. A fogyasztók már akkoriban is kételkedtek abban, hogy vajon a róluk megszerzett információkat miképp használják fel a vállalatok és ez napjainkra még több problémát vet fel, hiszen ma már sokkal több adatot lehet megszerezni az ügyfelekről (vagy akár a potenciális ügyfelekről) és ezen adatok nem legális (vagy minimum nem etikus) használatáról is egyre többet hallani, mindezek nyomán pedig a fogyasztókban (jogosan) csökkenhet a biztonság érzet és érezhetik magukat egyre sebezhetőbbnek a vállalatokkal szemben. (Swinhoe, 2021)

Ezeket az információval való visszaélés miatt kialakuló negatív érzeteket a szaknyelv „céginformációs kockázati együtthatónak” (firm information risk coefficient) nevezi. Akcura és Srinisavan (2005) kutatásának része volt egy kérdőíves kutatás is, mely során kiderült, hogy a megkérdezett egyetemi hallgatók 33%-a felfüggesztene minden továbbá vásárlást adott X vállalatától, továbbá 43%-uk valamilyen fajta extra előnyt várna, amennyiben a vállalat több keresztértékesítési kampánnyal célozná őket. Ebből jól látszik, milyen fontos, hogy előre meg tudjuk határozni, kik azok, akiket célozni érdemes és, hogy ténylegesen mely termékekkel tesszük ezt. Az is lényeges (ahogy a kutatásból is kiderült), hogy ténylegesen úgy érezzék a keresztértékesítéssel megkeresett fogyasztók, hogy meg vannak jutalmazva, tehát nagyobb valószínűséggel kapunk értékes információkat vásárlóinkról, ha valamilyen árendedményt biztosítunk ezért számukra. Egy másik lehetséges megoldás az Amazon példája, ahol a fogyasztók kezében van az irányítás, hogy mennyi keresztértékesítéshez kapcsolódó ajánlást vagy értesítést akarnak kapni.

Knott meghatározását Li et al (2011) fejlesztette tovább, miszerint a megfelelő terméket a megfelelő vásárlónak kell értékesíteni. Li és kutató társai emellett kritikus kérdésként határozták meg azt is, hogy mikor értékesítünk, hiszen hiába tudjuk, hogy egy jelenleg 18 éves vásárló szeretne majd életbiztosítást, ha azt majd csak 35 évesen fogja akarni, akkor jelenleg nem valószínű, hogy pozitív hatást tudunk elérni nála. Ez egy olyan probléma, amelyen sok

vállalat átsiklik, Li kutatása alapján hiába van több területen magas stratégiai prioritással jelen a keresztértékesítés, ennek ellenére rengeteg több millió dolláros kampány termel veszteséget és általánosságban ezek a kampányok a megkeresetteknek mindösszesen a 2%-ánál érnek el vásárlást az ezt követő három hónapban.

Az elemzők gyakran alkalmazzák a kampányorientált keresztértékesítést, mely során az elérhető termékek tulajdonságainak és a fogyasztók demográfiai adatainak a segítségével igyekeznek meghatározni azokat a vásárlókat, akiknél profitot lehet elérni a keresztértékesítés segítségével és ez alapján rendezik sorba őket.

A keresztértékesítést az egyik legfontosabb CRM eszköznek nevezhető, amely segítségével erősebbre fűzhető a fogyasztók és a vállalat közötti kapcsolat. (Kamakura et al, 2003) Ennek oka az, hogy minél több terméket vásárol a fogyasztó ugyanattól a vállalattól, annál magasabbá válik a váltási költsége a fogyasztónak, ezalatt a vállalat is egyre több információt szerez a fogyasztóról, annak vásárlási szokásairól és valószínűsíthetően emiatt egyre pontosabb ajánlatokkal tudja célozni. Mindezek mellett az is kiderült, hogy vevőélettartam értéke és a fogyasztói lojalitás is folyamatosan növekszik a sikeres keresztértékesítések után. (Kamakura et al, 2003)

Ezek alapján már jól láthatjuk, hogy a keresztértékesítés viszonylag komplex témakör, Schmitz et al (2014) Li és társaihoz hasonlóan megállapították, hogy a keresztértékesítési kampányok jelenleg egyáltalán nem működnek effektíven. A kutatásukból kiderült, hogy például a mérnöki és a kémiai szektorokban a keresztértékesítési potenciálnak mindösszesen az 1/3-át használják ki. Kitértem rá, hogy nem célozhatunk egyszerre minden fogyasztót, de fontos a kampány kialakítása során figyelmet szentelni arra is, hogy azokat a fogyasztókat is elérjük, akik nyitottak a termékajánlataink felé. Schmitz et al kutatása alapján ezen 1/3-os arány mögött legnagyobb részben az áll (58%), hogy az értékesítési munkatársak nem ismerik fel a lehetőséget és utánuk következően (26%) az értékesítési menedzserek inkompetenciája okozza ezt az arányt.

Ezzel kapcsolatosan viszont érdekes megállapítás, hogy az is kiderült, hogy nő a keresztértékesítési arány akkor, ha az értékesítési munkatársaknak nagyobb szabadságot adunk arra vonatkozóan, hogy eldöntsék, mely fogyasztókat és milyen termékekkel célozzák.

Az elektronikus adattárolás és rendszerezés fejlődésének hála, új lehetőségek jöttek létre a keresztértékesítés területén és azon belül már sokkal pontosabban tudjuk prediktálni azt, hogy kinek milyen termékekre és szolgáltatásokra lesz lehetősége. (Kamakura, 2004) 2004-hez képest

ezen a területen is rengeteg fejlődés történt, így ez még aktuálisabb, mint valaha, és az általa ajánlott mikromarketing is sokkal megvalósíthatóbbá, olcsóbbá vált a digitális marketing előretörésével. Azonban ő is felhívja arra a figyelmet, amelyre már a fogyasztói lojalitás területén is tettem említést, hogy a múltbeli vásárlások nem minden esetben határozzák meg azt, hogy milyen vásárlások várhatóak a jövőben, a trendek változhatnak ezért kritikus szemmel kell kezelni ezt a keresztértékesítés területén is és úgy felállítani a modelleket.

A nemzetközi szakirodalomban a felülértékesítés erősen alulreprezentált a keresztértékesítéssel kapcsolatosan, amelyet sajnálatosan én is meg tudok erősíteni ezen kutatásom alapján. Ez előrevetíti azt, hogy a felülértékesítés esetén minimum azok a hibák előfordulnak, mint a keresztértékesítés esetén, de valószínűleg a hiányzó, ismert legjobb módszerek (best practice-k) nélkül még nagyobb százalékban fordulhatnak elő problémák. Johnson és Friend (2014) a keresztértékesítéssel és a felülértékesítéssel kapcsolatosanközös pontként megállapították, hogy amennyiben erre a két területre fókuszál egy vállalat (a vevő megtartás mellett), akkor tízszer akkora növekedést érhet el a vevői kiadásokat (share of wallet) tekintve, mintha egyedül a vevő megtartásra fókuszálna. (Johnson és Friend, 2014)

Mindezek mellett kutatásukból kiderült az is, hogy a felülértékesítés segítségével az értékesítési munkatársak könnyebben kielégíthetik a magasabb minőségű termékek, illetve szolgáltatások iránti vásárlói igényeket. Az értékesítők motiváltabbak a felülértékesítést illetően, amennyiben ők maguk is úgy gondolják, hogy a prémium termékek megérik az árkülönbözetet, ezért fontos szempont, mint a fogyasztói lojalitás növelésének esetén is, hogy a dolgozóink ismerjék a termékeinket / szolgáltatásainkat és lássák annak előnyeit, hátrányait, hogy eldönthessék, hogy az adott vásárlónak érdemes-e terméket ajánlani és, ha igen, akkor melyiket.

II.3.2. Telekommunikációs és pénzügyi szektorban

A keresztértékesítés és a felülértékesítés esetén is megvizsgáltam, hogy milyen egyezések és sajátosságok fedezhetők fel a többi szektorhoz képest. Guillet (2020) megállapította, hogy az életbiztosítás szektorában az ügyfeleknek közel felében potenciál rejtőzik arra, hogy drágább, prémium szolgáltatást adhassanak el nekik és ezen szolgáltatások összesen mintegy 25%-kal növelhetik meg a szolgáltatásból származó bevételeket.

A következő példa a kínai telekommunikációs szektorra vonatkozik, amelyet Li (2009) kutatásában tett közzé. Li megállapította, hogy a telekommunikációs szektorban tevékenykedő vállalatok esetében rengeteg különböző adatuk képződik a fogyasztókról, beleértve az egyszerű

demográfiai adatokat, de a legkülönbözőbb viselkedési adatokat is, amelyek a különböző szolgáltatások használata során jönnek létre. A kínai vállalatok is meglátták a lehetőséget ezekben az adatokban, ezért már 2009-ben különböző üzleti intelligencia rendszereket építettek ki, amelyek segítségével komplex elemzéseket hajthattak végre támogatva ezzel többek között a keresztértékesítést is.

Egy másik kutatás a dél-koreai telekommunikációs szektort vizsgálta. Ahn et al (2010) megállapította, hogy a telekommunikációs szektoron belül a keresztértékesítés egy kifejezetten kritikus pont, amelyen szükséges a vállalatoknak jól teljesíteni, amennyiben növelni akarják a bevételeiket. Ennek oka az, hogy mivel az egyik legfejlettebb országról van szó és gyakorlatilag mindenkinek van előfizetése, ezért az előfizetők számát nagyon nehéz növelni és nagy a verseny a különböző szolgáltatók között, emiatt különböző hozzáadott szolgáltatásokkal kívánnak nagyobb bevételekre szert tenni. Ezen hozzáadott szolgáltatásokat a szaknyelv érték hozzáadó szolgáltatásnak (value added service-nek), röviden VAS-nak nevezi. Ezen szolgáltatások lehetnek különböző mobilon elérhető játékok, csengőhangok, kuponok vagy éppen telefonon keresztül történő tranzakciók.

A VAS szolgáltatások alapvetően azért lettek olyan népszerűek Dél-Koreában mert szüksége volt / van a telekommunikációs vállalatoknak kiegészítő szolgáltatásokra ahhoz, hogy csökkentsék az óriási költségekkel járó versenyt a hangalapú kommunikációs piacon. Azonban annak ellenére, hogy relatív széles termékpalettája van a szolgáltatóknak a VAS-on belül és ezáltal szinte minden korosztálynak tudnának valamilyen szolgáltatást ajánlani, sokan nem ismerik ezeket a szolgáltatásokat. Itt emelkedik ki az értékesítési dolgozók munkája: az ő dolguk megtalálni azokat a fogyasztókat, akik nyitottak egy vagy több kiegészítő szolgáltatásra, és meg kell határozniuk azt is, hogy mely szolgáltatások felelnének meg legtokéletesebben ezen fogyasztók igényeinek. Összegezvén Dél-Koreában a telekommunikációs szektorban a keresztértékesítés nem csak egy plusz bevételi forrás, hanem a piacon való megmaradás, a túlélés záloga. (Ahn et al, 2010)

Utolsó példám a banki szektorból származik, amely a keresztértékesítés és összességében a marketing történelem egyik legnagyobb botránya is, a Wells Fargo bank keresztértékesítési botránya. Eredetileg a vállalat vezérigazgatója (aki 2013-ban az év bankárja is lett) papíron nem állt rosszul hozzá a keresztértékesítéshez, elmondása szerint: *„Ahhoz, hogy sikeres legyél a keresztértékesítésben ezer dolgot kell jól csinálnod. Rengeteg kitartásra van szükség, sok pénzt kell költened a rendszerekre, az oktatásra, a dolgozók megfelelő*

ösztönzésére, elismerésre és időt kell szánnod arra, hogy megismered az ügyfeleid pénzügyi céljait”. Ez lehetne akár az összefoglalója is ennek a résznek, azonban a gyakorlatban John Stumpf nem úgy tett, ahogy mondta. (Tayan, 2019)

2013-ban kiderült, hogy az értékesítők agresszívan léptek fel annak érdekében, hogy elérjék a hozzájuk rendelt keresztértékesítési kvótákat, ez idő tájt 30 dolgozónak felmondtak amiatt, hogy az ügyfelek tudta nélkül kötöttek hozzájuk új termékeket. Tették ezt amiatt, mert folyamatos stressz alatt álltak, ha nem teljesítette egy bankfiók a napi célokat, akkor az hozzá adódott automatikusan a következő napihoz, ezzel még nagyobb feszültséget generálva. Korábban említésre került, hogy ahhoz, hogy effektíven tudjon értékesíteni egy értékesítő, ahhoz neki is hinnie kell a termékben és az ügyfél érdekeit kell szem előtt tartani, de itt ez egyértelműen megbukott.

A számokat tekintve először 2,6 millió dollárt térített vissza ügyfeleinek a bank és összesen 5,300 dolgozótól váltak meg, de itt még nem volt vége. John Stumpf vezérigazgatónak az amerikai szenátus előtt is meg kellett jelennie. Ezután Stumpfot felszólította a vezetőség, hogy adjon fel 41 milliónyi fizetségét, amely az egyik legnagyobb vezérigazgatói fizetés visszavonás volt a történelemben és a valaha volt legnagyobb a pénzügyi szektor területén. 2017-re a helyzet odáig súlyosbodott, hogy a FED korlátozást adott ki miszerint a Wells Fargo eszközmérete nem nőhet 1,95 trillió dollár fölé, ami egy példa nélküli döntés volt. Ezután összességében a bank még több mint 2 milliárd dollárt fizetett ki különböző perek következtében.

Ezen példa jól mutatja, hogy egy papíron jó célokkal elindított programban is benne van a kisiklás lehetősége, ezzel pedig óriási veszteségeket okozva a vállalatnak, nem csak pénzügyileg, hanem megbízhatóságát tekintve is.

II.4. Fogyasztómegtartás és lemorzsolódás

II.4.1. Meghatározása és jelentősége

Fogyasztó megtartás alatt alapvetően azt a folyamatot értjük melynek eredményeként a fogyasztók nem pártolnak el vállalatunktól, továbbra is termékeink / szolgáltatásaink igénybe vevői maradnak. A lemorzsolódás ennek gyakorlatilag az ellentétje, mikor valamilyen oknál fogva a fogyasztók kilépnek az ügyfélkörünkől és (legalábbis egy ideig) nem veszik igénybe termékeinket / szolgáltatásainkat.

Egyre fontosabbá válik meghatározni, mely ügyfeleink azok, akik potenciálisan elvándorolhatnak, és közülük is meghatározni, hogy kik azok, akiket érdemes megtartani (hiszen, ahogy lojalitást sem érdemes kialakítani azon vásárlóknál, akik veszteséget okoznak, úgy megtartani sem érdemes őket) és kik azok, akiket nem. Annak érdekében, hogy megtarthassuk értékes ügyfeleinket fontos, hogy az értékesítés után is figyelmet fordítsunk rájuk („post-sales”), ez magában foglalja a fogyasztókkal kapcsolatos elégedettséget és panaszvizsgálatát, amellyel egyrészt az adott fogyasztó esetében is csökkenthetjük a lemorzsolódás esélyét, illetve a tapasztalatokat felhasználva a jövőben más fogyasztók esetén akár a konfliktus kialakulását is meg tudjuk előzni. (Heidrich, 2016)

Heidrich-hez hasonlóan Veres (2017) is kifejtette, hogy nem minden ügyfelet érdemes (vagy legalábbis nem ugyanakkora erőfeszítésekkel) megtartani, megállapítása szerint ki kell számolni az ügyfélértékét a fogyasztóinknak és ez alapján kell dönteni arról, hogy mit érdemes megtenni azért, hogy továbbra is az ügyfeleink maradjanak.

Veres emellett kiemeli a bizalom szerepét: fontos, hogy a vásárlóink bízzanak bennünk, hiszen ez nagyban növeli a megtartás esélyét. Hasonlóan gondolkodik Bernschütz et al (2016) is, akik a bizalom kialakulását a vásárlói elégedettséghez köti, úgy véli az elégedettségen át vezet az út a bizalomhoz, a bizalom a megtartáshoz, a megtartás pedig segít a fogyasztói lojalitás kialakulásában. A megtartás eszközeiként említi a kilépési korlátokat (melyre már mi is korábban kitértünk váltási költségek néven), a különböző hűségprogramokat (mint a fogyasztói lojalitás során említett törzsvásárlói mérföldek) és összességében az ezeket (is) alkalmazó kapcsolati marketinget.

Bernschütz-el Poór (2017) is egyetért abban, hogy az ügyfél-elégedettség egy fontos tényező, amely az ügyfélmegtartáshoz vezet. Habár Poór művében a tanácsadás iparágával foglalkozott, összességében is helytálló lehet azon megállapítása, hogy a frontemberek (adott esetben a tanácsadók) olyan teljesítményt nyújtsanak, amellyel a fogyasztó elégedett lesz, sőt, úgy érzi, hogy vele különlegesen jól bánnak az adott helyen. A keresztértékesítés során megállapított legjobb módszer, miszerint úgy kell gondolkozni, hogy a fogyasztónak jó legyen, a fogyasztói megtartás növelésében is fontos tényező lehet.

A lemorzsolódás előrejelzését, mint a fogyasztói magatartáselemzés egy fontos részeként definiálták (Eichinger et al, 2006). A fogyasztók megtartása különösen fontos tényező a telekommunikációs szektorban, mivel egy új fogyasztó megszerzése körülbelül öt-nyolcszor akkora költségekkel jár, mint megtartani egy jelenlegit.

Habár tökéletes előrejelzést nem lehet felállítani és mindig lesznek fals pozitívok, fontos, hogy igyekezzük minél pontosabban meghatározni, hogy mely ügyfeleinknél jelenhet meg a lemorzsolódás kockázata, mivel amennyiben ezt sikeresen tesszük, jelentősen növekedhetnek vállalatunk bevételei.

II.4.2. A telekommunikációs szektorban

Mutanen et al (2006) Hwang tanulmányai alapján felhívja a figyelmet arra, hogy a fogyasztók lemorzsolódása a legnagyobb probléma a mobilszolgáltatóknál jelenleg a piacon lévő rendkívül nagy verseny miatt. Vereshez hasonlóan kifejti, hogy az ügyfélértéket figyelembe kell venni, mivel amennyiben egy ügyfél korábban lemorzsolódik, mint az átlag (vagy mint ameddig gondosabb kapcsolat marketinggel maradt volna), akkor a vevőélettartam értéke is csökkenni fog és ezáltal a vállalat bevételei is.

Reichheld et al (1990) az telekommunikációs iparág vizsgálatát követően már 1989-ben megállapította, hogy a lemorzsolódási ráta 5%-os csökkentésével már 85%-os profitnövekedés lenne elérhető, amely a fogyasztói lojalitás növekedéséhez hasonló és százalékosan kimagaslóan nagy érték. (Bhatti et al, 2016) Egy másik tanulmány levezeti, hogy a fogyasztók egyre nyereségesebbé válnak a vállalat számára (például a hozzájuk kötődő marketing költségek az idő előrehaladtával csökkenthetőek, míg egy-egy új vásárló esetén az akvizíciós költségek kimagaslók). (Darzi – Bhat, 2017) Érzékelhető tehát, hogy az említett szektorban rendkívüli értékkel rendelkezik a fogyasztói megtartás, mégis viszonylag kevés kutatás készült még ezen a területen.

A megtartási arány a szolgáltatói iparágakban alacsony, amennyiben a kiszolgáló személyzet nem elég képzett és ezáltal nem teljesíti a vásárló elvárásait, amely egybecseng a Poór által a tanácsadás területén megállapított gondolatokkal. Rávilágítottak, hogy a kulcsemberek a jó kapcsolat kialakításában a kiszolgáló személyzet, hiszen velük találkozik a legtöbbet a fogyasztó, őket tudja adott esetben azonosítani a vállalattal. Általában nem elég a fogyasztói elégedettség, más változók is szerepet játszanak abban, hogy az ügyfél a vállalatnál marad-e vagy sem, a szerzők szerint ezek a pozitív szójuttatás (WOM), az újvásárlási kedv és a vállalathoz való lojalitás az elégedettség mellett. Ezen változók esetében észre kell venni, hogy nem állnak egymástól külön, hiszen az elégedettség könnyen indíthat el pozitív reklámot, illetve hozhatja meg az újvásárlási kedvet, amelyek összességéből hosszú távon kialakulhat a fogyasztói lojalitás is. (Darzi – Bhat, 2017)

II.4.3. A pénzügyi szektorban

A telekommunikációs szektor után a bankszektor vizsgálgva érdemes kitérni Cohen et al (2006) munkásságára, akik az új-zélandi bankok területén vizsgálták a fogyasztó megtartást és annak hatásait. Megállapították, hogy a pénzügyi intézetek vezetőinek a szemében magas prioritású problémának kell lennie a fogyasztói lemorzsolódásnak, és meg kell állapítaniuk, mik lehetnek azok a változók, amelyek ezen lemorzsolódáshoz vezetnek. A kutatók a következő változókat emelték ki, amelyek a banki szektor esetén kiemelten fontosak lehetnek a fogyasztó megtartással kapcsolatosan: a dolgozók teljesítményét, felkészültségét, kedvességét (ahogy már a fentebb többször is szó esett erről), a pénzügyi intézetek elhelyezkedését, szabályait, illetve díjszabását.

Rendkívül érdekes és kiemelendő megállapítás, hogy saját kutatásuk alapján az új-zélandi bankszektorban a legfontosabb változónak a fogyasztói elégedettséget, a vállalati imázst és a váltási korlátokat emelték ki, amely gyakorlatilag megegyezik a korábban a fogyasztói lojalitás és a keresztértékesítés esetén észleltekkel. A demográfiai adatokat vizsgálva megállapították a szerzők azt is, hogy általánosságban a magasabb iskolázottsággal nőnek az elvárások is a bankok felé, amellyel a vállalatoknak számolniuk kell. A fiatalok sokkal kisebb eséllyel maradnak hosszabb távon egy banknál (46,5%), mint a 61 évnél idősebbek (95%), amelynek oka többek között az lehet, hogy a keresztértékesítési kampányok miatt az idősebbek már sokkal jobban kötődnek egy-egy bankhoz, mint fiatalabb társaik. A kutatásból kiderült, hogy a nem vagy éppen a háztartások átlagjövedelme kevésbé jelentősen befolyásolja a fogyasztó megtartást. (Cohen et al, 2006)

A fogyasztó megtartása egy hosszú folyamat eredménye, ahol alapvetően az ügyfelek a múltbéli évek alatt kialakult tapasztalataik alapján döntenek el, hogy érdemes-e az adott vállalat ügyfelének maradni. (Liu és Wu, 2007)

Tajvani kutatásukból kiderült, hogy a fogyasztói megtartási ráta növelését érdemes a keresztértékesítéssel ötvözni. Megvilágították, hogy az országban a különböző pénzügyi intézetek ezt olyan módon teszik, hogy különböző szövetségekbe állnak a bankok biztosítókkal és tőkebefektetőkkel, ezáltal képesek egy rendkívül nagy portfóliót lefedni, értékes terméket ajánlani és tenni mindezt úgy, hogy felkészült szakemberek és kiszolgáló személyzet állnak a vállalat és az ügyfelek rendelkezésére. Emellett ebben a szektorban a vállalat megítélése nagyon fontos tényező, hiszen a fogyasztók nem akarják a befektetéseiket olyan vállalatra bízni, akiben a közvélemény nem bíz meg valamilyen oknál fogva. Amennyiben a közvélemény, a vállalati

imázs jó, az könnyen növelheti a vállalatba vetett bizalmat, ami a fogyasztói megtartási ráta növekedéséhez vezet hosszútávon. Kiemelték továbbá az új-zélandi kutatáshoz hasonlóan, hogy a bankfiókoknak is nagy szerepük van a fogyasztói megtartás növelésében, fontos ugyanis, hogy ezeken a helyszíneken kényelmesen tudják ügyeiket intézni a fogyasztók, megbízható és szakértő munkatársak közreműködésével. (Liu és Wu, 2007)

Utolsó példám ebben a témakörben Kassim és Souiden (2007) kutatása akik az Egyesült Arab Emírségekben vizsgálták a bankszektoron belüli fogyasztói lemorzsolódást. Kutatásukban megállapították, hogy a fogyasztói megtartás vagy éppen a lojalitás mérése a szolgáltatói iparágban sokkal nehezebb feladat, mint a gyártói iparágakban, mivel kevésbé pontosan meghatározható, hogy a fogyasztó miért választ egy-egy szolgáltatást, illetve a szolgáltatás értékelése is nagyban függ attól, hogy éppen az adott pillanatban milyen érzelmi állapotban van egyrészt a fogyasztó másrészt pedig a dolgozó, akivel kapcsolatba lép. Az írók a kutatás végére megállapították, hogy az Egyesült Arab Emírségekben a bankszektoron belül a fogyasztói lemorzsolódás tekintetében a szolgáltatás minősége nem fontos befolyásoló tényező, ami más kutatásokkal is egybevág. Azonban, szintén más általunk említett kutatáshoz hasonlóan a vállalati imázst, mint fontos befolyásoló tényezőt említette meg az elégedettség mellett.

Az olyan szektorokban, ahol gyakorlatilag a nagyvállalatok már nem tudják a fogyasztói létszámot növelni (ami a telekommunikációra és a pénzügyi szektorra is alapvetően igaz), vagy csak nagyon költséges módon, ott különösen nagy szerepet kap az a felfogás, hogy a legértékesebb meglévő ügyfeleket a lehető legtöbb ideig megtartsák, mivel sokkal nagyobb mértékben növelhető ezáltal a profit, mint új ügyfelek szerzésével. (Hadden et al, 2007)

Hadden és társai a lemorzsolódó ügyfeleket először két felé bontották: a nem önkéntesen lemorzsolódókra és az önkéntesen lemorzsolódókra. A nem önkéntesen lemorzsolódók közé tehetőek azok az ügyfelek, akikkel a vállalat szakította meg a kapcsolatot (például azért, mert megszegtek bizonyos szabályokat), az önkéntes lemorzsolódók közé pedig azok, akik úgy döntöttek, hogy felfüggesztik a szolgáltatás igénybevételét vagy a termékek megvásárlását. Az önkéntes lemorzsolódást további két csoportra osztották: nem szándékos és szándékos lemorzsolódókra.

Nem szándékos lemorzsolódáson azokat az ügyfeleket értik, akiknek valamilyen oknál fogva nem volt szükségük továbbra a szolgáltatásokra (például a pénzügyi helyzetük miatt nem tudták megengedni, más életkorba kerültek, vagy éppen elköltöztek egy olyan országba, ahol

nem elérhető a szolgáltatás). Általánosságban a nem szándékosan lemorzsolódók az önkéntes lemorzsolódóknak csak egy alacsony hányadát teszik ki. A nagyobb hányadot kitevő szándékosan lemorzsolódók azok, akik közül alapvetően lehetőség van megtartani a számunkra értékeseket. A szándékos lemorzsolódás oka lehet például az, hogy a versenytársaknál találkozott egy jobb ajánlattal, vagy éppen az, hogy nem volt megelégedve a vállalatunk által nyújtott szolgáltatásokkal.

Ezen példák is jól szemléltetik, hogy mint a banki, mind a szolgáltatói szektorokban egyre nagyobb komolysággal kezelik a vállalatok a fogyasztói megtartást. A telekommunikációs piacon, ahol az éves lemorzsolódási ráta 20 és 40% között mozog világszinten, felmerül a probléma, hogy ha a vállalatok pótolni akarják a kiesett ügyfeleket, rengeteg pénzt kell minden évben ügyfélakvizícióra és marketingre költeniük, amennyiben nem sikerül a lemorzsolódási rátát csökkenteniük. (Ahn et al, 2006)

Összességében érzékelhető, hogy a defenzív marketingstratégia egyre inkább felértékelődik, hiszen a vállalatok nagyobb profitnövekedést érhetnek el azzal, ha sikerül a fogyasztómegtartást növelniük. Hatásosabb tehát erre összpontosítani, mint a versenytársaktól való ügyfélszerzésre összpontosítani, amikor megvan annak a veszélye, hogy ez idő alatt mi is fogyasztókat veszünk felőlük. (Ahn et al, 2006)

II.5. Adatbázis-marketing

II.5.1. Adatbázis marketing

Keller és Kotler (2016) szerint ahhoz, hogy a vállalat kellő betekintést nyerjen az ügyfeleikre és azok vásárlási szokásaikra, szükséges az adatokat egy előre meghatározott módszer alapján gyűjteni, rendezni és elemezni. Keller és Kotler definíciója szerint a „...*vevői adatbázis az egyes vevőkre és potenciális vásárlókra vonatkozó átfogó adatok olyan szervezett gyűjteménye, amely aktuális, elérhető és felhasználható például az értékesítési lehetőségek előkészítésére, az értékesítési lehetőség értékelésére, termék vagy szolgáltatás eladására vagy vevőkapcsolatok fenntartására. Az adatbázis-marketing a vásárlói és más adatbázisok (termék, beszállítói, viszonteladói) létrehozását, fenntartását és kapcsolatteremtésre, üzletkötésre és vevőkapcsolatok kialakítására való felhasználását jelenti.*”. (Keller – Kotler, 2016)

Ennél lényegesen rövidebben határozza meg ezt Malhotra és Simon (2017) (Gyulavári et al (2017) szintén ezt a meghatározást használja), miszerint az adatbázis marketing „*azt jelenti, hogy számítógéppel rögzítik és követik nyomon az ügyfelek lényeges jellemzőit és vásárlásaik részleteit*”.

Az adatbázis megfelelő alkalmazásával elérhetőek olyan direkt marketing kampányok, amelyek esetén a válaszadási arány a 10%-os szintet is elérheti. Keller - Kotler, (2016) Malhotra és Simon (2017) hasonló következtetésre jutottak, miszerint az adatbázis marketing alapján olyan pontos szegmentációt végezhetünk, amelynek a segítségével sokkal pontosabb és eredményesebb direkt marketing kampányokat végezhetünk.

Az adatbázis-marketing és a direkt marketing közötti kapcsolatra Gyulavári et al (2017) is felhívták a figyelmet, utalva arra, hogy az adatbázis marketing kialakulása összeköthető a direkt marketing fejlődésével, mint egy azt segítő eszköz.

Az adatbázis-marketing felhasználási területei azonban a direkt marketingen belül és azon kívül is rendkívül sokrétűek lehetnek: segítségével hatékonyabban oszthatjuk el az erőforrásainkat (legyen szó akár pénzügyi, időbeli vagy humán erőforrásról), versenyelőnyt építhetünk ki az ezt nem használó versenytársainkkal szemben, (hiszen effektívebben és gyorsabban dolgozhatunk), hibákat előzhetünk meg (a rendszer például automatikus üzeneteket küldhet, ha nem olyan terméket vagy szolgáltatást akarunk nyújtani egy ügyfélnek, amelyért fizet /fizetett). (Gyulavári et al, 2017) (Keller – Kotler, 2016)

Ezekon felül a direkt marketingen belül az adatbázis marketing segíthet növelni a vevőmegtartási rátát és lehetőséget nyújt a lojalitás növelésére és a kereszt- és felülrértékesítés során összepárosítani a megfelelő termékeket a hozzá tartozó vevőkkel és időpontokkal, továbbá a már lemorzsolódott vevőink visszaszerzésében is eszközként szolgálhat egy adott vállalatnak. (Keller – Kotler, 2016)

II.5.2. Adatbányászat

Ahogy a fentiekben elhangzott az adatbázis marketing alkalmazása alatt nagy mennyiségű adat alapján kell különböző megállapításokat tenni, ám ezek az adatok általában rendezettek és emiatt könnyebben is hasznosíthatók. Azonban vannak olyan adatok, mint például a vásárlók viselkedési adatai, amelyeket egyáltalán nem ilyen egyszerű gyűjteni vagy kezelni, vagy amikor valamilyen összefüggéseket akarunk megtalálni adat és adat között például egy pontosabb direkt marketing vagy mikromarketing kampány esetén. Ekkor jön segítségünkre az adatbányászat. Gyulavári et al (2017) szerint az adatbányászat *„több, mint egyszerű adatlekérdezésből, összesítésekkel és alapstatisztikai vizsgálatokkal történő információszerzés.”*, emellett fontos előnyeként említi a gyorsaságot és a pontos objektív információk elérését. (Gyulavári et al, 2017)

Az adatbányászati tanácsadással foglalkozó vállalatok jelentőségét emelte ki Veres et al (2017), hiszen, mivel sok esetben az adatokkal rendelkező vállalatoknak nincsenek megfelelői dolgozói vagy infrastruktúrája arra, hogy elvégezze ezeket az elemzéseket, ezért külső vállalatokat kérnek fel erre. Példának felhozta az általam is említett telekommunikációs illetve pénzügyi szolgáltató vállalatokat, ahol az adattudomány segítségével sokkal pontosabb szegmentáció végezhető, mint előtte. (Bose-Malhapatra, 2001)

Az adatbányászat alapú marketing segítségével elvégzett projektek rendkívül nyereségesek tudnak lenni, ezt igazolják a Microsoft alapján elmondottak is, miszerint a magatartás alapú direkt marketing segítségével 76%-kal növelhető azok száma, akik egy adott reklámra kattintanak. (Keller- Kotler, 2016)

Az adatbányászat azonban egy bonyolult, több lépéses feladat, a nagy profit növekedés kecsegtető, de ezért több lépést is szükséges megtenni. Mindenekelőtt szükség van adatokra, amely egy, de általában több különböző adatbázisból származik, melyeket össze kell kapcsolni, meg kell tisztítani (mely az adatelemzés legnagyobb részét képezi, gondoljunk csak például arra, ha elírások keletkeztek), illetve elő kell készíteni olyan módon, hogy az elemzéshez használni tudjuk. Fontos, hogy eldöntsük mely attribútumokra lesz szükségünk, továbbá, hogy egyáltalán mely attribútum pontosan mit jelent (adott esetben akár az is előfordulhat, hogy két adatbázisban ugyanazon attribútum név alatt más fajta, más aggregálású adatok találhatók, ezzel számolni kell). (Bose – Mahapatra, 2001)

II.5.3. Mesterséges intelligencia

Meg kell említeni a mesterséges intelligencia jelenlétét, hiszen felvirágzása óta nagyban segíti az adatbányászatot és az adattudósok munkáját és ezzel indirekt az adatbázis marketing használatát is. Danyi et al (2020) a mesterséges intelligenciához két különböző definíciót határozott meg, az egyik szerint: *„Minden olyan algoritmus mesterséges intelligenciának tekinthető, amely az emberi döntéseket helyettesíti.”*, a másik szerint pedig *„Minden olyan algoritmus mesterséges intelligenciának tekinthető, amely képes a kombinatorikus robbanásból eredő exponenciális problémák polinomiálissá való visszavezetésére.”* (Danyi et al, (2020)

A mesterséges intelligencia segítségével egy adott vállalat profitjai akár 10-20%-kal is növekedhetnek, mivel a gépi tanuló algoritmikus segítségével egy vállalat meghatározhatja, hogy egy adott vásárlónak az adott pillanatban milyen árat érdemes mutatni, úgy, hogy közben figyeli a versenytársak árait, a potenciális ügyfél viselkedését vagy gyakorlatilag bármit. A jövőben az sem tűnik lehetetlennek, hogy a gépi tanuló algoritmusok képesek legyenek

meghatározni az ügyfelek rezervációs árát, amellyel általában maguk a fogyasztók sincsenek tisztában. (Danyi et al, (2020)

Kacsukné – Kiss (2019) könyvükben a mesterséges intelligenciát Danyiék első definíciójához hasonlóan fogalmazták meg: szerintük a mesterséges intelligencia a *„számítógépek használata olyan feladatokra, amelyekhez magas szintű emberi intelligencia, azaz tanulási, gondolkodási és szimbólumkezelési képesség szükséges.”*

A gépi tanuló algoritmusok segítségével olyan állandó, könnyen reprodukálható tudás kerülhet a vállalatok birtokába, amely korábban lehetetlen volt, hiszen az emberek elfelejtenek dolgokat, nem mindig teljesítenek ugyanúgy, pihennek, a köztük lévő tudásátadás is lassú, sőt néha szinte lehetetlen, viszont ez a gépek esetében egyáltalán nem így van. (Kacsukné – Kiss, 2019)

A gépi tanuló algoritmusok sem képesek magukban megváltani a világot. Sikeres működésükhöz először létre kell őket hozni, figyelni működésüket, fejleszteni őket. (Kacsukné – Kiss, 2019) Emellett azt sem szabad elfelejteni, hogy a mesterséges intelligencia alapvetően nem az ember helyett dönt, hanem segít a döntésekben, hiszen egy hibrid duó (ember + mesterséges intelligencia) 85%-kal produktívabb tud lenni, mint egy ember – ember vagy egy gép – gép páros. (Wharton, 2018)

A mesterséges intelligencia fontosságát tovább növeli az is, hogy Danyi et al (2020) szerint az árazás a közgazdaságtan egyik, ha nem a legbonyolultabb területe és a mesterséges intelligencia nagy segítségére van a marketingnek abban, hogy megállapíthassa az árakat, a potenciális ügyfeleket, az ajánlott terméket, és azt, hogy mikor áruljunk, így láthatjuk azt is, hogy a korábban említett témakörökhöz is szorosan kapcsolódik.

II.5.4. A telekommunikációs és pénzügyi szektorban

Végezetül néhány gyakorlati példa az adatbázis marketingre vonatkozóan: a már korábban említett és hírhedt Wells Fargo bank képes volt az összes tranzakció adatát elemezni különböző demográfiai adatok segítségével és ennek a segítségével olyan direkt marketing kampányokat tudott létrehozni, ahol az egy főre jutó termékek száma elérte a 4,7 darabot, ami több, mint kétszerese a 2,2-es bankiparban jelen lévő átlagnak. (Keller – Kotler, 2016)

Szintén a bankiparra jellemző példa a mesterséges intelligencia segítségével történő kockázatelemzés, hitelemzés vagy éppen a csődelőrejelzés. Ez utóbbit például úgynevezett

neurális hálós megoldásokkal akár 97%-os pontossággal lehet előre jelezni és ezáltal el lehet kerülni a különböző esetlegesen előforduló veszteségeket. (Bose- Mahapatra, 2001)

Egy másik kiváló példa a Qwest nevű telekommunikációs céghez kötődik, akik a vállalati adatbázisukat éves szinten kétszer alaposan átkutatták annak érdekében, hogy felmérjék az ügyfelek esetében azt, hogy körülbelül mekkora összegeket fognak elkölteni mind telefonálásra mind összességében, ezzel összhangban pedig maximalizálták az egyes ügyfelekre fordítandó marketingkeretet, hiszen tudták mely ügyfelek esetében van akkora a várható nyereség, hogy érdemes egyedileg is célozni. (Keller – Kotler, 2016)

A telekommunikációs szektorban azonban már a Qwest előtt jóval elkezdték felfedezni a mesterséges intelligenciában rejlő lehetőségeket. Az első feljegyzett felhasználásig egészen 1988-ig kell visszamenni, azonban akkor még nem marketing célokra használták, hanem különböző komplex rendszerek diagnosztikájára. Nem sokkal később, 1993-ban már a telekommunikációs vállalatok működésének egyik szükséges alapjaként tekintettek a mesterséges intelligenciára. (Qi et al, 2007)

Később az évek alatt, mint mesterséges intelligenciai használati terület a telekommunikáció piacán előtérbe került a lemorzsolódás vizsgálat, a hálózati terheltség előrejelzés vagy éppen a csaló hívások kiszűrése. A British Telecom a Netmap nevű rendszer segítségével szűrte ki a hívók közül azokat, akik a többi fogyasztót kívánták különböző módokon átverni, és habár ez alapvetően nem marketing, hanem mint biztonsági lépés lényeges, ettől függetlenül fontos jelezni, hogy nagyban növelheti a vállalat jó hírnevét, ami, ahogy korábban is felmerült, fontos befolyásoló tényező a marketing területén is. (Bose – Mahapatra, 2001)

Azonban nem mindig van esély ilyen sikeres megoldásokat létrehozni. Az adatbázis marketinggel kapcsolatosan is vannak problémák, amelyekre figyelni kell: egyrészt nem mindig érdemes vevői adatbázist létrehozni (például rendkívül ritkán vásárolt termékek esetén, alacsony CLV esetén vagy akkor, amikor nagyon költséges információkat gyűjteni), másrészt a vevői adatbázisok bevezetése után a cégek 70%-a nagyon alacsony vagy semmilyen javulást nem tapasztalt például a félresikerült tervezés vagy a túl magas költségek miatt. Ezeken kívül lényeges továbbá az is, hogy a vezetés egyhangúan kiálljon a vevői adatbázis kialakítása, és üzemeltetése mellett és a dolgozók is tisztában legyenek az előnyökkel, legyen meg a kellő belső támogatás. Azonban mindezek jelenléte esetén is megtörténhet az, hogy a vásárlók ismerik a saját értéküket, és a hűségükért alapvetően elvárnak bizonyos engedményeket,

szolgáltatásokat vagy egyszerűen nem örülnek, nem engedik a megkereséseket, és ezáltal a nyereségnövekedés sem érhető el, legalábbis nem a várt mértékben. (Keller – Kotler, 2016)

II.6. A kereszt- és felülrértékesítés előrejelzésére alkalmazott prediktív módszerek bemutatása

Ebben a fejezetben három különböző módszert ismertetek. Elsőként kitérek az NPTB módszerre, amely a szakirodalmakban és a gyakorlatban nem kifejezetten ismert, majd bemutatom a dolgozatom gerincét adó és a kvantitatív kutatásomban is alkalmazott véletlen erdő gépi tanulási módszert, végezetül a marketingben sokat használt logisztikus regresszió kerül bemutatásra. Ezen három módszerre esett a választásom, mivel be akartam mutatni a marketingben általában használt, a jövőben alkalmazandó és egy véleményem szerint méltatlanul elfeledett módszert is.

II.6.1. NPTB

Már a keresztértékesítés témája során előtérbe került Knott (2002) illetve hozzá kapcsolódóan az NPTB modell említés szintjén, mely ebben a fejezetben kifejtésre kerül

Az NPTB modell jelentése magyarul „következő megvásárlandó terméket” (next-product-to-buy) jelent, és alapvető célja a keresztértékesítés hatékonyságának növelése. Nevéből adódóan az NPTB modell célja, hogy megjósolja, mely terméket fogja várhatóan egy bizonyos fogyasztó a legnagyobb valószínűséggel megvásárolni. Fontos leszögezni, hogy az NPTB modell nem köti ki, hogy pontosan milyen előrejelző módszer kerül alkalmazásra, így használható hozzá logisztikus regresszió vagy neurális háló is, ezen felül arról is az adattudósnak kell döntenie, hogy mely adatokat fogja pontosan felhasználni az előrejelzéséhez és majd milyen módon fogja az eredményeket tesztelni. Ebből adódóan nincs kizárólag egy konkrét NPTB modell, hanem gyakorlatilag végtelen számú kombináció létezik, emiatt is pontosabb előrejelzési egyenletnek nevezni az NPTB-t, mint konkrét megoldásnak. (Knott et al, 2002)

Az NPTB négy fő lépésből áll: Először is szükséges összegyűjtenünk, rendeznünk és előkészítenünk az adatokat, másodszor ki kell választanunk a statisztikai modellek széles skálájából egyet (vagy akár többet és az eredményeket összehasonlíthatjuk kiválasztva a legjobb megoldást), a modell sikeres lefutása után alkalmazhatjuk a későbbiekben részletezett NPTB egyenlőséget, majd végül a modell eredményei alapján kell pontozni az ügyfeleket és megcélózni őket a marketing kampányunkkal.

Az NPTB elemzés során használt képlet (logisztikus regresszió esetén) a következő:

Annak a valószínűsége, hogy i fogyasztó következő megvásárolt terméke j lesz

$$= \frac{1}{1 + e^{-V_{ij}}} \quad (1)$$

$$V_{ij} = \sum_{k=1}^K \beta_{jk} X_{ijk} \quad (2)$$

3. ábra NPTB elemzés során használt képlet.

Forrás: Knott, 2002

Ezen egyenletben az X_{ijk} és a K értékek előrejelző változók minden i fogyasztó esetén és azt állapítják meg, hogy a fogyasztó vásárolni fog-e legközelebb j terméket. β_{jk} a statisztikai modell által kiszámolt koefficiensek. J termék esetén összesen J különböző logisztikus regressziós modellre van szükségünk. A függő változó minden modell esetén 0 és 1, azaz vagy megvásárolja az ügyfél j terméket, vagy nem. Független változóink az X -ek.

Az NPTB modell esetén további érdekes megfigyelt jelenség lehet a termék affinitás, azaz annak a valószínűsége, hogy x fogyasztó mekkora eséllyel vásárolja meg a terméket, akkor, ha k termék már megvan neki.

A kampány esetén a fogyasztói célcsoport kiválasztásakor egy különösebb elemzést nem igénylő, egyszerű megoldás lehet az, hogy meghatározunk egy kampány büdztét, sorba rendezzük a fogyasztókat a vásárlási valószínűség szerint és a top 20%-ot keressük meg. A legjobban viszont az NPTB sikerességét a kampányok ROI-ja mutatja: az NPTB direkt marketing kampány esetén a ROI 530,1% volt, miközben a heurisztika és a vásárolt adatbázis csoport is negatív (-16,8 és -30,2%-os) értékeket ért el. Ezen példa alapján egyértelműen látszik, hogy sokkal jobban sikerült célozni a terméket a fogyasztóknak ezzel, mint a régi megszokott heurisztikán alapú megoldással.

További fontos észrevétel volt, hogy az NPTB modell a saját kontroll csoportját is lényegesen felülmúlta: a kiválasztottak esetében a válaszadási ráta 1,13% volt míg a kontroll csoport esetében 0,50%, azaz a modell alapvetően olyan fogyasztókat talált, akik nem vették volna meg más csatornán az adott terméket, ellenben a heurisztikán alapuló megoldással, ahol a megkeresettek válaszadási rátája 0,44% volt, a kontroll csoporté pedig 0,42% azaz ez esetben magas volt a kannibalizáció, hisz a fogyasztók mindenképpen megvásárolták volna a terméket.

Kutatása során Knott a gyakorlatban is letesztelte az NPTB modellt, egy banki hitel termék értékesítése során. A teszt fázis során 6 különböző csoport lett elkülönítve: az első

csoport egy a bank esetében szokásos heurisztikán alapú levél alapú direkt marketinggel megkeresett csoport volt (olyan fogyasztókat kerestek, akinek a háza 100,000 dollárnál többet ért). A második csoport szintén levélben lett megkeresve, de az NPTB módszer alapján kerültek kiválasztásra, abban az esetben, ha az adott termék az első vagy a második legvalószínűbb termék volt számukra. A harmadik csoport olyan fogyasztókat tartalmazott, akik nem voltak a bank jelenlegi ügyfelei (tehát egy vásárolt adatbázisból származtak) és szintén levélben lettek megkeresve. A negyedik, ötödik és hatodik csoport pedig kontroll csoportok voltak a három előbbi módszerből származóan, akik nem lettek levélben megkeresve. Utóbbi kontroll csoportok beiktatására azért volt szükség, hogy mérni tudják, mekkora kannibalizációt okoznak a különböző megoldások.

A kutatás eredményei rendkívül sikeresek voltak: Az NPTB direkt marketinggel megkeresettek esetén a válaszadási ráta 1,13% volt a heurisztikás megkeresettek 0,44%-hoz és a vásárolt adatbázis 0,10%-hoz képest, továbbá a megkeresett fogyasztókra eső bevétel 93,28 \$ /fő volt a 29,63 és a 22,21 \$ /főhöz képest.

Összességében a kutatásban kiderült, hogy az NPTB tökéletes módszer lehet a fogyasztó megtartás növelésére, illetve a különböző keresztértékesítési kampányokon elért profitok növelésére. Az előrejelzés alatt Knott (2002) különböző statisztikai modelleket is használt, melyek között kifejezetten nagy különbséget nem vélt felfedezni. Azonban visszatérő esemény volt, hogy a legfontosabb előrejelző változó egyértelműen az, hogy jelenleg milyen termékei / szolgáltatásai vannak a fogyasztóknak.

A mintavételhez kapcsolandó megállapítás, hogy lehet véletlen vagy rétegzett mintavételről is szó, utóbbira akkor lehet szükség, ha termékpalettánkon nagyon sok népszerű elem van, mert ekkor a véletlenszerű mintavétel esetén könnyen elfogult lehet a modell.

Az NPTB modell esetében azonban két probléma is fennáll: egyrészt az NPTB modellek csak azt képesek meghatározni, hogy mi lesz a fogyasztó következő vásárlása, azt nem, hogy ez pontosan mikor fog megtörténni. Másrészt a modell nem veszi figyelembe azt, hogy milyen sorrendbe vásárolt korábban különböző termékeket a fogyasztó, azaz nem tud felfedezni egy olyan mintát, hogy például egyre drágább termékeket vásárol. Ezek olyan problémák mellyel modellalkotás közben számolni kell és lehetőség szerint megoldást találni rájuk. (Prinzie- van den Poel (2007)

II.6.2. Véletlen erdő (Random Forest)

A véletlen erdő módszer bemutatása előtt szükségesnek tartom röviden bemutatni a döntési fa módszert is, hisz, ahogy lentebb láthatóvá válik majd, a véletlen erdők gyakorlatilag döntési fákat alkalmaznak. Már a döntési fákat is mint hatékony, problémafüggetlen és jól skálázható statisztikai módszernek nevezte a szakirodalom, hiszen attól függetlenül, hogy viszonylag egyszerűen értelmezhető, egyszerű módszerről van szó, könnyen működőképes tud lenni nagyobb adathalmazok esetén is. Azonban a módszer nem volt problémamentes, például amennyiben a betanító adatsor esetén nagy szórás volt tapasztalható, abban az esetben az előrejelzés könnyen félresiklik. (Olaru-Wehenekel, 2003)

A döntési fák esetén további előnynek számít, hogy relatív könnyen értelmezhetőek és jeleníthetőek meg grafikusan is (nem csak táblázattal vagy számokkal szemléltethető). Továbbá maga a modell „gondolkodása” is hasonló az emberi gondolkodáshoz, amely segítségével sokszor képes pontos becslésekre.

Definíció szerint a döntési fa alatt egy olyan (digitális) fát értünk, melynek minden csomópontja egy attribútumot jelent, az elágazások különböző döntési helyzeteket és minden levél egy eredményt. A döntési fa mögött rejlő algoritmusra több különböző megoldás is létezik (például a CART, amely a híres Gini indexet használja), nincsen kifejezett „legjobb” módszer. Maga a modell gyakorlati alkalmazását szintén többféleképpen oldhatjuk meg, például különböző programozási nyelvek (Python, R) vagy grafikus statisztikai szoftverek (SPSS, SAS, KNIME, Rapidminer) segítségével. (Sanjeevi, 2017)

A döntési fával kapcsolatosan fontos kitérni a fa különböző elemeinek a nevére, amelyre gyakorlatban és szakirodalomban is gyakorta hivatkoznak. A döntési fa gyökere egy attribútum, amely esetén a hozzá kapcsolódó szabály segítségével két vagy több halmazra osztódik az adathalmazunk (itt válik először több részre az adathalmaz). Ágnak nevezzük egy teljes döntési fának egy részét, vágás („splitting”) alatt azt értjük, amikor egy csomópont a hozzá kapcsolódó szabály alapján két vagy több részre bomlik. Levélnek hívjuk a döntési fának azon csomópontjait, ahol már nem osztódik több részre, nem alakul ki új szabály, azaz a modell abba az irányba befejeződik. Ezeken felül a döntési fánkon „metszést” is végezhetünk, amikor valamilyen tudományos megfontolás alapján eltávolítunk a fáról az alcsomópontokat. (Chowdary, 2020)

A döntési fa modellek továbbfejlesztését, a véletlen erdő azaz random forest modellt Leo Breiman alkotta meg 2001-ben és a következőképpen definiálta: „*A véletlen erdő egy olyan*

osztályozó algoritmus, amely osztályozó fák összessége: $\{h(x, \Theta_k), k=1, \dots, \text{ahol } \{\Theta_k\} \text{ független, azonos elosztású véletlen vektorokat jelöl és mindegyik fa egy egységnyi eredményt (szavazatot) ad a legnépszerűbb osztályra az adott } x \text{ bemeneten.}\}$ (Breiman, 2001, 2. old.)

Breiman nem kötötte ki, hogy pontosan hány darab fára van szükség ahhoz, hogy a lehető legjobb eredményt elérjük, ezzel kapcsolatban Oshiro et al (2012) végeztek kutatásokat és arra a következtetésre jutottak, hogy 29 különböző adathalmaz esetén 64 darab fa fölött nem történt nagy változás a pontosságban, és 128 fa fölött pedig szinte semmilyen, így a véletlen erdő módszeren alapuló algoritmusok esetén célszerű lehet egy 64 és 128 fa közötti értéket megcélózni, és ennek a segítségével úgy kaphatunk pontos eredményeket, hogy a modell is képes gyorsan, kevesebb erőforrás befektetésével lefutni.

A véletlen erdő módszer sikerességét mutatja, hogy napjaink egyik legsikeresebb jelenleg elérhető előrejelző algoritmus. Sokrétű felhasználhatósági és relatív egyszerű beállításának hála használták 3D-s tárgyak felismerésére, levegőminőség előrejelzésre és a bioinformatikában is a teljesség igénye nélkül. Összességében elmondható, hogy nagy sikere többek között annak köszönhető, hogy már kis mintaméret esetén is jól tud működni. A fenti előnyök miatt egyesek a modern idők legsikeresebb általános felhasználású algoritmusának nevezték és a Google vezető közgazdásza is kiállt emellett. (Biau – Scornet, 2016)

A sikeresség oka még ezen felül, hogy a különböző döntési fák alapvetően korrelálatlanok egymással, így a fák által létrejött n darab megoldás dönti el szavazás útján, hogy melyik az általuk leghelyesebbnek ítélt eredmény és ezzel kiküszöbölhetőek az egyes döntési fák hibái. (Yiu, 2019b)

Annak érdekében, hogy elérjük, hogy korrelálatlanok legyenek egymással a döntési fák, két feltételnek kell eleget tennie az algoritmusunknak: Az egyik az úgy nevezett „bagging eljárás”, azaz miszerint meg kell határoznunk, hogy hány elemű legyen a tréning adatsorunk és minden egyes fának ennyi adatot kell adnunk ugyanazon adatok közül, téve mindezt úgy, hogy ismétlődő helyettesítéseket használunk. Tehát például amennyiben a tréning adataink [1,4,6,8], akkor az egyik fánknek a [8,8,4,1] adatokat adhatjuk meg például, azaz ugyanúgy négy elemű a tréning adatsor, de nem helyezkedik el benne az összes adat, és van benne ismétlődés. (Yiu, 2019b)

A másik feltétel, amelynek eleget kell tennünk az az, hogy az attribútumok (jellemzők) esetén a megadottak közül minden fa csak egy részét választhatja ki, tehát amennyiben az attribútumaink [kék-lila-zöld-sárga], akkor az egyik fánk megkaphatja a [kék-lila], egy másik

pedig a [zöld-sárga] attribútumokat. Ezzel és a tréning adatok felosztásával pedig képesek leszünk elérni, hogy korrelálatlanok legyenek egymással a döntési fáink és ezáltal pontosabb, hibáktól védettebb adatokat kaphatunk. (Yiu, 2019b)

II.6.3. Logisztikus regresszió

A harmadik bemutatott statisztikai modell a döntési fa és a véletlen erdő után a logisztikus regresszió. A logisztikus regresszió a döntési fákhöz és véletlen erdőkhez hasonlóan egy osztályozó algoritmus, mely esetben kategorikus függő változóval dolgozunk. Tehát a mi esetünkben például, amennyiben a függő változónk az, hogy x fogyasztó megvásárolja-e y terméket z időben, akkor egy kategorikus változóval van dolgunk, hiszen az eredmény vagy nem (0) vagy igen (1). (Ehhez a példához hasonló eseteket, amikor csak két változó van bináris logisztikus regresszióknak is nevezik, míg több változó esetén multinominális logisztikus regresszióról van szó.) (Agrawal, 2017)

A logisztikus regresszió gyakran használt módszer, mivel kategorikus függő változók esetén egyszerű, gyors és hatásos megoldást kínál az előrejelzési problémákra. Azonban amennyiben túl sok kategorikus változót használunk fel az algoritmus során, azt nem képes jól kezelni, illetve a nemlineáris változókat át kell transzformálnunk lineáris változókká, hogy használhassuk az algoritmushoz.

A logisztikus regressziós modell egyenlete relatív egyszerű: $Z = \beta_0 + \beta_1 x_1$, ahol $Z = \log(a \text{ függő változó valószínűsége})$. Fontos még kitérni a költség funkcióra: a logisztikus regresszió a költség funkciót használja arra a célra, hogy büntesse saját magát akkor, ha rossz eredményt jósolt. A logaritmus regresszió esetén a költség funkció a következőképpen számítható ki: ha a tényleges kimenetel 1, akkor a költség = $-\log(\text{jósolt_valószínűség})$, ha a tényleges kimenetel 0, akkor viszont a költség = $-\log(1-\text{jósolt_valószínűség})$. Ezeket a költségeket minden kimenetel esetén kiszámítva és összeadva megkaphatjuk a regresszió teljes költségét. (Yiu, 2019a)

III. A kutatás kérdései és módszere

III.1. Kutatási célok

Kutatásom központi kérdése azzal foglalkozik, hogy: *A véletlen erdő módszer segítségével milyen módon és mekkora mértékben segíthető elő a hazai telekommunikációs szektorban a felül- illetve keresztértékesítés?*

A kutatásom célja a kérdés részletes körbejárása: központi elemeként szerepel egy hazai telekommunikációs vállalat valós adatain végzett felülrértékesítési előrejelzés - véletlen erdő módszer segítségével. Célom egy olyan keretrendszer és „útmutató” felállítása is egyben, amely segítségével a tanulmány elolvasása után az olvasók képesek lesznek saját vállalataikban is hasonló minőségű véletlen erdő modelleket létrehozni. Ennek érdekében a problémát az adatok importálásától egészen az eredmények értékelésig körbejárom, kitérve arra, hogy a modellalkotási folyamatok egyes lépéseinél pontosan milyen okból választottam bizonyos módszereket és mely más módszereket vettem alapul.

Annak érdekében, hogy a modell megvalósítása vonzóbb legyen a marketingmenedzserek számára egy esettel is készültem, amelyben a saját modellem segítségével mutatom be, hogy egy magyarországi nagyvállalat esetében milyen hatása lehet egy ilyen modellnek egy vállalat pénzügyi teljesítményére. Ezután kitérek még a modell által generált nem vagy kevésbé számszerűsíthető várt hatásokra, mint például arra, hogyan hat a fogyasztói élményre, vagy dolgozóink munkájára. Végezetül kitérek az általam észlelt legnagyobb kihívásokra, potenciális veszélyekre és a modell hiányosságaira is.

A vállalati adatbázison végzett modellépítés esetében abból az üzleti problémából indulunk ki, hogy jelenleg az ügyfélszolgálat otthoni internet felülrértékesítéshez köthető hívásainak mindösszesen 26,16%-a sikeres (és még így is az otthoni internet felülrértékesítés esetében a legmagasabb ez az arány), és nem is találunk meg minden olyan ügyfelet, aki magasabb internetcsomagra váltana. A modell célja, hogy növelje az értékesítések számát és a sikeres hívások arányát is. Azért esett a választásom az internet migrációra, mert általánosságban elmondható, hogy a modellek rosszabbul működnek kiegyensúlyozatlan adatbázisok esetén, tehát ott, ahol nem egyenlő arányban oszlanak meg a kategóriák (jelen esetben az, hogy valaki migrált vagy sem). Ezen a téren az internet migrációnak volt a legmagasabb sikerességi aránya (az összes hívás 11,77%-ánál történt internethez kapcsolódó felfelé migráció), ezért hüvelykujj szabályként ezt választottam.

Más problémákra is alkalmazható és építhető modell az általam végigvezetett lépések útján, azonban a kiegyensúlyozatlan adatok kezelése több időt vehet igénybe, illetve a modell valamelyest rosszabbul teljesíthet például egy vezetékes telefonszolgáltatáshoz köthető keresztértékesítési feladat esetén (mivel itt alulmintázás helyett felülmintázásra lesz szükségünk, amely torzíthatja az eredményeket).

Hipotézisek alkotására nem került sor, mivel a marketingben megszokott kutatási módszerekkel ellentétben a gépi tanuló algoritmusok esetén ez nem szükséges. A gépi tanuló algoritmusok nem egy-egy változó / jellemző hatásait vizsgálják bizonyos problémára, hanem komplex problémákat kerülnek úgy, hogy egyszerre többszáz vagy akár több ezer jellemzőt és az ő kapcsolatukat vizsgálják. Az én modellemben is több, mint 180 jellemzőt és ezek kapcsolatát vizsgálom, valamilyen szinten megjelenik egyes jellemzők fontossága (egy adott jellemző mennyire számít erős vágási kritériumnak), amelyre a modell értékelésekor ki is fogok térni.

Az értékesítési előrejelzés elsődleges célja az értékesítések és a sikeres hívások arányának növelése, ezáltal az összes hívás szám csökkentése. A modell előnye, hogy nem egy-egy jellemzőt ragad ki, hanem a való életben is jelen lévő problémát modellez le (amelyben még a modell által ismerteknél is több jellemző szerepel). Azzal, hogy létrehozunk egy modellt, képesek leszünk az értékesítés komplex kérdéseit elemezni és javítani tudunk vállalatunk pénzügyi eredményein. A közismertebb megoldások esetében a költséges átalakítások végén kiderülhet, hogy hiába volt bizonyos változóknak nagy hatása az értékesítésre, az jórészt függött más változóktól, amely kapcsolatot nem vagy csak részben voltunk képesek értelmezni és ezáltal a projekt végén nem a várakozásunknak megfelelő eredményeket tapasztaltunk. A legerősebb és leggyengébb vágó kritériumok bemutatása során több olyan példát is találhatunk, amelyekkel kapcsolatban szakirodalmak vagy tapasztalat alapján nem feltétlen végeznénk kutatást, vagy pont, hogy feleslegesen nagy erőket fordítanánk rá.

III.2. A vizsgált adatbázis

A kutatásom egy magyar telekommunikációs cég adatain alapul. A vizsgált adatbázis összesen kb. 42 millió adatponttal rendelkezik, melynek 10%-át teszik ki a hiányzó értékek. Az adatbázis összesen 182 jellemzővel (a mesterséges intelligencia, illetve a gépi tanulás területén nem változónak nevezzük az „oszlopokat”) rendelkezik, amelyek egy-egy adott híváshoz kapcsolódóan rengeteg extra információt hordoznak.

Ez a kutatást mindenképpen elősegítette, hiszen rengeteg területre vonatkozóan álltak rendelkezésre adatok, amelyekből a későbbiekben lehetőségem volt szelektálni, amely egy kevésbé időigényes folyamat annál, ha különböző csoportokból, különböző rendszerekből kellett volna adatokat szerezni és összesíteni őket egy adatbázisban.

A teljesség igénye nélkül az adatbázisban lévő jellemzők tartalmazzak különböző demográfiai adatokat, a meglévő szolgáltatásokra vonatkozó adatokat, különböző direkt

marketinghez kapcsolatos információkat, hibabejelentéseket, Facebook üzeneteket vagy akár azt, hogy volt-e tartozása, vagy korábban kereste-e már az ügyfélszolgálatot, és ha igen, akkor milyen ügyben.

Az adatok a vállalati rendszerből történő exportálás után egy CSV fájlban kerültek tárolásra.

III.3. Az elemzéshez használt program és módszer

Az elemzésemhez a magyar kutatási életben kevésbé ismert KNIME Analytics Platform nevű programot használtam, mely egy térítésmentes open-source alkalmazás. A programra azért esett a választásom, mivel korábbi kutatásom során már használtam adattisztítási célokra, továbbá a data science és machine learning platformok közül a KNIME az egyik vezetőnek számít, nem csak tanulónak, hanem mindenki számára teljesen ingyenes.⁹

A platform használata könnyen megtanulható, „fogd és vidd” („drag and drop”) rendszerrel működik, ahol különböző csomópontokat (node-okat) használhatunk fel a projektünkhöz, kapcsolhatjuk össze őket különböző módokon és állíthatunk be paramétereket.

A KNIME előnye, hogy mivel nyílt forráskódú (open-source¹⁰) szoftver, ezért rengeteg különböző bővítmény (extension) létezik hozzá: integrálhatunk Java Script, Python, R programozási nyelvből különböző kódokat, előre létrehozott csomópontokat is hozhatunk be szintén a harmadik felektől, vagy akár létrehozhatjuk a saját csomópontjainkat is. A különböző általunk elkészített munkafolyamatokat (workflow-okat, tehát magukat a projektünkhöz használt csomópontjainkat és azok kapcsolatát) feltölthetjük a KNIME oldalára, megoszthatjuk másokkal és integrálhatjuk másukét a saját projektjeinkbe, saját adataink használatával, amellyel jelentős időt lehet megtakarítani.¹¹ Mindezek mellett, amennyiben valamilyen problémába ütközünk fejlesztésünk során, a hivatalos fórumon is gyorsan segítséget kaphatunk, sokszor KNIME alkalmazottak által, akik ezen a platformon keresztül térítésmentesen segítenek.¹²

A KNIME egy nemzetközileg elismert szoftver több különböző szektorban is. Sikeresen hasznosították a hadászatban, az autó-gyógyszeriparban, de a kutatásom szempontjából

⁹ <https://www.knime.com/about/news/knime-recognized-by-gartner-as-a-leader-in-data-science-and-machine-learning-platforms-2019>

¹⁰ <https://www.knime.com/open-for-innovation>

¹¹ <https://hub.knime.com/>

¹² <https://forum.knime.com/>

kiemelten fontos pénzügyi és telekommunikációs szektorban is olyan vezető vállalatok, mint a Deutsche Telekom, a Daimler vagy éppen a Siemens.¹³

Habár a KNIME Analytics Platform ingyenes, bizonyos extra, különböző (nagy) vállalatoknak fontos funkciók fizetős csomagokhoz lehetnek kötve, ehhez a KNIME három fajta konstrukciót is kínál KNIME Server néven, ennek használata által a fogyasztók például magasabb titkosításban kezelhetik az adataikat, illetve rendkívül kapacitásigényes Big Data munkafolyamatokat futtathatnak távolról.¹⁴

Végül, de nem utolsó sorban egyik nagy előnye a KNIME Analytics Platformnak, hogy az adatbetöltés, az adattisztítás, az előrejelzés, a tesztelés, a vizualizáció és a kiértékelés is elvégezhető benne, anélkül, hogy elhagynánk a programot.¹⁵

Az elemzéshez használt módszer, amelyet a KNIME segítségével alkalmazni fogok a véletlen erdő lesz. A véletlen erdő, ahogy a szakirodalmi elemzés során említettem, egy gépi tanuló algoritmus, ami döntési fák összességéből áll és ezen döntési fák szavazatai alapján kategorizál a modell.

A véletlen erdő modell kitűnően alkalmazható kategorizálási feladatokhoz, mint amilyenem nekem is dolgom volt. A modell előnyei tartozik, hogy:

- a hiperparaméterek (azon paraméterek, amelyek alapján a modell tanul és nekünk kell megadnunk) közül kevés megadására, optimalizálásra van szükség¹⁶
- jól megbirkózik a magas dimenzionalitású adatbázissal: több mint 180 jellemző van, ezeket egyszerre kezelni rendkívül bonyolult, azonban ahelyett, hogy az egész modellből vennénk el jellemzőket, az egyes fák maguk állítják sorba a jellemzőket, illetve az elméleti összefoglalóban bemutatott módon kisebb mintán dolgoznak.¹⁷
- csökkenti a túltanulás esélyét azzal, hogy több fa véleményére támaszkodik egyszerre¹⁸

¹³ <https://www.knime.com/solutions>

¹⁴ <https://www.knime.com/knime-software/knime-server-pricing>

¹⁵ <https://www.knime.com/>

¹⁶ <https://builtin.com/data-science/random-forest-algorithm>

¹⁷ <https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706>

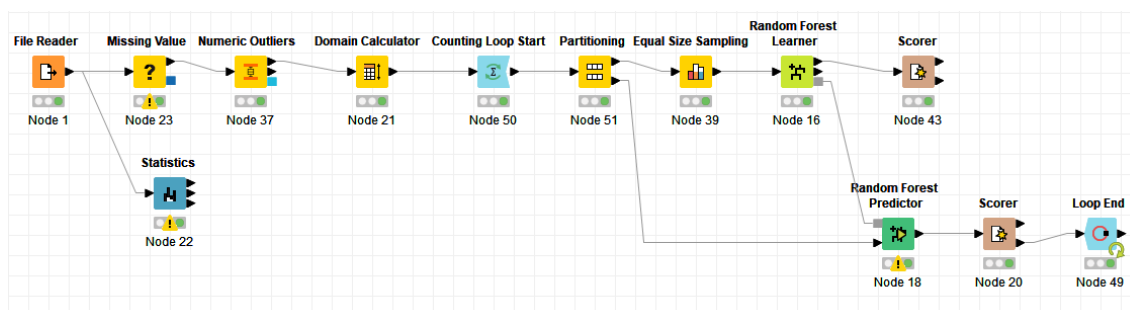
¹⁸ <https://towardsdatascience.com/random-forest-models-why-are-they-better-than-single-decision-trees-70494c29ccd1>

A modellnek az előnyei mellett hátrányai vannak, amelyeket a modellezés során számba kellett vennem:

- A modell eredményeinek a mélyére ásni, hogy pontosan mi miért lett az, ami, önmagában nehéz, a modell fekete dobozként funkcionál⁹. Ahogy fentebb említettem a döntési fák előnye, hogy relatív könnyen értelmezhető a logikájuk, azonban mikor már egy erdőnyt veszünk belőlük, akkor ez sokkal nehezebbé válik. Ennek ellenére az eredmények után bemutatom a legjelentősebb és legkevésbé jelentős választott vágási kritériumokat, ezzel elősegítve a megértést.
- Sok fa esetén nagy számítási kapacitást és adott esetben sok időt igényel a modell lefuttatása. Ennek kiküszöbölésére a szakirodalom alapján felmértem az ideális erdő méretét, ennek elérése után pedig saját laptopomon 2 perc alatt képes volt a tanító és a tesztelő része is lefutni a modellnek. Nagyvállalatok esetében a felhőben való futtatással ez az idő még tovább csökkenthető, vagy éppen rendkívül magas adatmennyiség esetén tartható.¹⁹

IV. Kvantitatív kutatás:

IV.1. KNIME folyamat ábra:



4. ábra A véletlen erdő modellhez tartozó folyamat ábra. Forrás: Saját szerkesztés.

A fenti ábra a modell munkafolyamatát mutatja be, rajta az összes a végső változatban szereplő csomóponttal és a köztük lévő kapcsolatokkal. A továbbiakban a kutatás lépéseit bemutatva a fenti csomópontok értelmet nyernek, mint a kutatás és a modellépítés részei. Azonban két csomópont csak technikai célokra szolgál, ezért ezeket itt mutatom be röviden. A „statistic node” a különböző jellemzőkről ad nekünk statisztikai eredményeket, mint például minimum, maximum vagy variancia. Ez a különböző adattípusok beállításához jelentett nekem segítséget. A „domain calculator node”-ra pedig azért volt szükség, mert az adattípusok

¹⁹ <https://www.mygreatlearning.com/blog/random-forest-algorithm/>

átállítása után a programon belül elvesznek bizonyos jellemzőkhöz rendelt értékek, mint a felvehető minimum és maximum értékek, ezeket ennek segítségével kell pótolni.

IV.2. Adatok importálása és előkészítése

Az előrejelző folyamat első fontos lépése (az adatok kinyerése után) az adatok importálása, amelyet a KNIME-on belül a korábban ismertetett csomópontok segítségével tehetünk meg. A „File Reader” képes az adatok gyors beolvasására, képes felismerni az oszlop elválasztókat, illetve a fejléceket is. Az adatbeolvasás idejét tekintve összehasonlítható azzal, ha a CSV fájlt a Microsoft Excelben kívánjuk megnyitni.

A következő lépés a megfelelő adattípusok beállítása, amelyet szintén a „File Reader-ben” tehetünk meg. Ez a beállítás kritikusan fontos, hiszen a „File Reader” alapból igyekszik megtippelni, hogy milyen adattípusa lehet a különböző jellemzőknek, de sokszor nem jár sikerrel, ezért az ajánlásait ellenőrizni kell.

Az adatbázis kódolását tekintve szinte kizárólag csak számokat tartalmazott, emiatt a program numerikusnak minősítette szinte az összes jellemzőt, ami azonban nem volt jó. Példa erre a város vagy a kor, ahol amennyiben számként kezeljük őket, akkor a tanuló algoritmus úgy fogja érzékelni, hogy a 2-es kódolású város (pl. Pécs) egy fele akkora értékű adatpont mint a 4-es kódolású társa (pl. Szeged), pedig nem erről van szó. A két város teljesen más, egymástól különbözőnek kell őket venni, nem határozható meg, hogy egyik kétszerese lenne a másiknak. Az életkor esetén pedig szintén arról beszélhetünk, hogy egy 10 és egy 20 éves között 10 év különbség van és egy 80 és egy 90 éves között is, azonban teljesen másként kell a tíz év különbséget a két esetben kezelni, ezért ezen jellemzőket szöveggé konvertáltam.

IV.3. Kiszűrt jellemzők:

Kiszűrésre kerültek adott értékesítési hívásnál történt egyéb eredmények, hiszen ezeket nem láthatja, nem prediktálhat belőle a modellt, mivel akkor elfogult eredményeket kaphatnánk. Abból kiindulva, hogy egy értékesítésnél csak egy terméket vesz / hosszabbít a vásárló, amennyiben látná, hogy bármely más esetben pozitív (1-es) a vásárló, akkor ez nagyban elősegíthetné az előrejelzési képességeit, de a való életben ezeket az adatokat még nem látnánk, hiszen csak egy hívás végén tudjuk leképezni, hogy mi az, ami sikeres felül és keresztértékesítés volt és mi az, ami nem. Ide tartozik még a hívás pontos idejét jelző jellemző is, hiszen ezt szintén nem láthatja előre a modell és csak egy adott hívás pillanatában képződik le.

Kivételre került a kliens ID is, mivel amennyiben ez bent maradna szintén téves megállapításokra juthatna a modell és az itt jelen lévő (alapvetően random) különbségekre

építene különböző előrejelzéseket. Hasonlóan kell eljárni a különböző technikai jellegű adatok esetén, amelyeket szintén nem kell és nem is szabad jellemzőként hasznosítani.

IV.4. Hiányzó értékek kezelése

A hiányzó értékek kezelése az adatelőkészítés egyik legfontosabb lépése. A hiányzó adatokat többféleképpen lehet vizsgálni és változtatni rajtuk, nincs kifejezett „legjobb módszer”, ezért sok különböző módszert kipróbáltam és kiértékeltem az eredményeket.

Itt és a későbbiekben is igaz az, hogy egy-egy módszer kipróbálása során úgynevezett ciklusokat („loop”-okat) helyeztem a modellbe, azaz ötször került lefuttatásra egy 64 fából álló véletlen erdő modell és az öt eredmény esetében a migrálók (azaz az '1'-es értékek) Pontosság és Felidézés értékeinek az átlagát hasonlítottam össze. A Pontosság és a Felidézés is 0-1 közötti értéket felvehető mutatók, amíg a Pontosság azt mutatja meg, hogy a modell mekkora arányban találta meg a ténylegesen migrálókat (azaz 1-eseket) addig a Felidézés azt mutatja, hogy az összes a modell által migrálónak jelölt fogyasztóhoz képest mekkora arányban vannak a ténylegesen migráló fogyasztók.

A könnyű összehasonlítás érdekében az eredményeket a Pontosság (Precision) érték alapján sorba rendeztem, majd pedig a legrosszabb Pontosságú értéket vettem a kiinduló 100%-os értéknek, mind a Pontosságú, mind a Felidézés (Recall) esetén. Fontos, hogy minden mérés és eredmény esetén csak az éppen vizsgált módszer módja változott, minden más változatlan maradt.

Fontos leszögezni, hogy a kutatás ezen szakaszában a Pontosság és a Felidézés mérése / javítása is egyaránt fontos volt számomra, pusztán egyszerűségi okokból választottam a Pontosság alapú rendezést.

Az összesített eredményeimet a numerikus értékek behelyettesítése esetén az alábbi táblázat mutatja:

	Pontosság	Felidézés
Mozgóátlag	103,84%	106,45%
Kerekített átlag	103,32%	109,82%
Átlag	103,24%	120,21%
„0” érték	103,07%	114,87%
„9 999 999” érték	102,77%	115,67%
Nem változtatunk	102,35%	112,16%
Minimum érték	101,86%	115,57%
Lineáris interpoláció	101,78%	107,10%
Medián	101,76%	112,31%
Maximum érték	101,01%	112,69%
X értéket megelőző érték	100,74%	106,44%
X értéket követő érték	100,50%	103,11%
Átlagos interpoláció	100,00%	100,00%

1. Táblázat A hiányzó értékek kezelésére használt módszerek (numerikus értékek esetén) összehasonlítása.

Forrás: Saját szerkesztés.

A fenti táblázaton jól látszódik, hogy habár a mozgóátlag rendelkezik a legjobb Pontossággal, a Felidézése relatív alacsony volt, azonban a legjobb Felidézéssel rendelkező „sima” (számtani) átlag is alig marad el a nála valamivel jobban teljesítő megoldásoktól, emiatt úgy döntöttem, hogy ezzel fogok a továbbiakban számolni.

Az átlag, mint hiányzó értékeket helyettesítő érték használata egyáltalán nem újkeletű dolog, hiszen amennyiben nincs lehetőségünk arra, hogy megszerezzük különböző más adatforrásokból a hiányzó értéket, egy jó becslést szerezhünk ezzel. A módszerrel kapcsolatban probléma lehet, ha egy-egy jellemzőn belül túl sok a hiányzó érték, az átlagok ezáltal felborítják a modell működését, de a mi esetünkben szerencsére ilyen probléma nem volt és a módszer a fenti táblázat alapján is jól működött.

	Pontosság	Felidézés
Átlag - X értéket követő érték	108,07%	153,50%
„9 999 999 érték” – módusz	100,70%	84,82%
Átlag – módusz	100,14%	80,17%
Átlag – nem változtatunk	100,46%	103,92%
„9 999 999 érték” – nem változtatunk	100,00%	100,00%

2. Táblázat A hiányzó értékek kezelésére használt módszerek (szöveges értékek esetén) összehasonlítása.

Forrás: Saját szerkesztés.

A fenti táblázatban az előzőhöz hasonlóan azt láthatjuk, hogy milyen eredmények születtek egyes módszerek esetén. A szöveges jellemzők kiváltására szóló módszerek az első oszlopban szereplő második értékek minden esetben, az első az előző táblázatban is használt és kiválasztott numerikus módszer, a használt átlag minden esetben jelen van és a szintén jó eredményeket hozó fix értékű behelyettesítéssel is méréseket végeztem két esetben. Az eredményeken jól látható, hogy habár amikor a szöveges értékek módusát használták a Pontosság valamelyest megnőtt, a Felidézés nagyon magas százalékban leesett, ezért ezeket az opciókat elvetettem.

Az is jól látszik, hogy a hiányzó értéket követő érték használata érezhetően magasabb Pontosságot és sokkal jobb Felidézést mutatott. Azonban kutatói megfontolásból, elemzésem során úgy döntöttem, hogy nem fogom ezt a megoldást választani, tekintettel arra, hogy nem magyarázható meg jelen adatbázis esetében, hogy miért pont a következő értékkel történne a behelyettesítés, hiába hoz ez jó eredményeket. Úgy döntöttem, hogy ezért meghagyom a hiányzó értékeket, és bár ez nagyon primitívnek hangozhat, azonban a véletlen erdő módszer képes a hiányzó értékeket is számításba venni és ez alapján prediktálni (kivéve akkor, ha az előre jelezhető értékben van hiányzó érték, azonban adatbázisunk esetén ilyen esetről nem volt szó). Fontos megjegyezni, hogy olyan értékek esetén, ahol ténylegesen nem létezik az adott információ a „nincs érték” maga a helyes érték, ezért is preferált ez a módszer. A gyengesége ezzel szemben az, hogy vannak olyan értékek, amelyek valahol bizonyosan léteznek, de az adatokat valamilyen ok miatt (elérhetetlen adatbázis, adatok helytelen gyűjtése, vállalati támogatás hiánya stb) nem lehet elérni. Gyakorlatban érdemes törekedni arra, hogy a vállalat által gyűjtött adatok közül minél többet megszerezzünk, ezzel pontosítva az előrejelzést.

IV.5. Kiugró értékek kezelése

A hiányzó adatok kezelése után a következő fontos feladat a kiugró értékek kezelése. A kiugró értékek esetén négy különböző paraméternek a változtatásával járó eredményeket

vizsgáltam. Az első a kiugró érték kezelés során felhasznált „R” érték. Az R értékek ahhoz szükségesek, hogy képesek legyünk meghatározni az adatok kvantiliseit, erre különböző módszerek léteznek ezeket különböztetik meg a különböző R értékek. (Az R értékek között hasonlóság, hogy jórészt az empirikus eloszlásfüggvényen és a lineáris interpoláción alapszik valamilyen formában.)

	Pontosság	Felidézés
93-7 R^2	103,44%	94,61%
93-7 R^4	102,91%	105,75%
93-7 R^8	102,82%	99,47%
93-7 R^9	102,65%	93,37%
93-7 R^5	101,97%	99,47%
93-7 R^6	101,18%	100,09%
93-7 R^7	100,87%	98,23%
93-7 R^3	100,58%	97,17%
93-7 R^1	100,00%	100,00%

3. Táblázat A kiugró értékek kezelésére során alkalmazott R értékek eredményeinek összehasonlítása.

Forrás: Saját szerkesztés.

A fenti táblázatból jól látszódik, hogy habár az R^2 érték hozza a legjobb Pontosságot, a Felidézés értéke a második legalacsonyabb az összes R érték közül, azonban az R^4 érték valamelyest alacsonyabb Pontosság érték mellett a legjobb Felidézés értéket hozza, ezért erre került a választásom. Az általunk választott R^4 érték az empirikus eloszlásfüggvény lineáris interpolációjaként állnak elő. Jelen esetben választásom pusztán az eredményekre alapszik, hiszen a modell pontosságára ez volt a legjobb hatással. Azt, hogy nincs kifejezett legjobb módszer az is mutatja, hogy különböző statisztikai szoftverekben (és például a KNIME-ban is) többféle módszerből választhatunk, a „győztes” kiválasztása a kutatóra van bízva.

A következő meghatározandó érték a kiugró értékek kezelése során a k érték volt. A „k” érték az interkvartilis terjedelem szorzóját jelenti, azaz, hogy az interkvartilis terjedem hányszorosa alatti, illetve feletti érték esetén vesszük kiugrónak az értékeket.

	Pontosság	Felidézés
k =1,5	102,49%	109,93%
k= 1,6	102,33%	104,41%
k= 1,4	100,91%	103,68%
k= 1,7	100,19%	103,95%
k= 1,3	100,00%	100,00%

4. Táblázat A kiugró értékek kezelése során alkalmazott k értékek eredményeinek összehasonlítása.

Forrás: Saját szerkesztés.

Nemzetközileg az 1,5-es érték van sztenderdnek véve, amelyet finomítani lehet kisebb hangolásokkal, ezért is néztem meg az 1,3-1,7 közötti értékeket, azonban, ahogy a táblázatból is látszódik Pontosság esetében is valamivel jobban teljesített a $k=1,6$ értéknél, a Felidézés esetén pedig jelentősebb mértékben a legjobb volt, így azt állapíthatjuk meg, hogy ezen érték kiválasztása esetén volt képes a legjobban szűrni az igazi kiugró értékeket és maximalizálni a predikciós eredményt.

Ezután azt kellett meghatároznom, hogy a kiugró értékekkel mi történjen:

	Pontosság	Felidézés
Csere az X értékhez legközelebbi értékre	100,07%	88,46%
Csere hiányzó értékre	100,00%	100,00%

5. Táblázat A kiugró értékek cseréje során alkalmazandó módszer kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

Itt viszonylag egyszerű dolgom volt mivel csak két lehetőség állt rendelkezésünkre. Egyrészt lecserélhettük a kiugró értékeket a hozzájuk legközelebbi eső értékre vagy pedig hiányzó értékke alakíthattuk őket. Ahogy a fenti eredményekből is láthatjuk a Pontosság esetén alig van különbség a Felidézés esetén azonban a hiányzó értékre való csere esetén sokkal jobb eredményt ad. A módszer hátránya, hogy ezáltal új hiányzó értékeink adódnak a tisztítás után is, azonban ahogy korábban említésre is került ezeket a hiányzó értékeket képes a modell tanulni és értelmezni, amennyiben azonban nem ilyen modellt használnánk, akkor akár rosszabb eredmények esetén is az X értékhez legközelebbi értékre való csere választása lenne a megfontolandó döntés.

A kiugró értékekre vonatkozóan az utolsó meghatározandó feladat a cserélni kívánt értékek kiválasztása:

	Pontosság	Felidézés
Összes	104,53%	84,40%
Csak a kiugróan magas értékek	103,23%	81,58%
Csak a kiugróan alacsony értékek	100,00%	100,00%

6. Táblázat A kiugró értékek során a cserélni kívánt értékek kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

A táblázatban jól látszik, hogy amennyiben csak a kiugróan alacsony értékeket kezeltük, abban az esetben a Felidézés erőteljesen növekedett, ez alapján könnyen választhatnánk ezt a megoldást a modell esetében. Azonban kutatói megfontolásból úgy döntöttem, hogy ezen pontra már helyesen meghatároztam azt, hogy milyen kritériumok alapján szűrjem ki a kiugró értékeket és amennyiben már sikerült jól meghatároznom mely értékek lesznek a kiugrók, az

esetben kezelni kell az extrém alacsony és az extrém magas eseteket is, hiszen hiába kaptam jelenleg jó eredményt, lehet, hogy a jövőben új ügyfelek, illetve ügyfélcsoportok esetén félre mérne a modell.

IV.6. Tanuló és teszt adatok szétválasztása

A következő fontos lépésként a meglévő adathalmazomat fel kellett bontani egy teszt és egy tréning adatsorra. Fontos, hogy ezeket már az adattisztítás után elkülönítsük, hiszen így elkerülhető az, hogy olyan adatokból tanuljon a modell, amelyeket nem szabadna látni, és így oldható meg laboratóriumi körülmények között is az éles tesztelés. A következő tábla az eredményeket mutatja (a számok az összes adatból kivett százalékos arányt mutatják: az első szám a tréning adatsorba sorolt adatok arányát a második pedig a teszt adatsorba soroltakét):

	Pontosság	Felidézés
93% tanító - 7% teszt	103,26%	108,46%
70% tanító – 30% teszt	102,04%	92,88%
60% tanító – 40% teszt	101,22%	85,31%
80% tanító – 20% teszt	101,17%	103,32%
75% tanító – 25% teszt	100,88%	99,01%
65% tanító – 35% teszt	100,75%	86,73%
90% tanító – 10% teszt	100,66%	100,65%
85% tanító – 15% teszt	100,00%	100,00%

7. Táblázat A megfelelő tanító-teszt arány kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

A fenti táblázatból meghatározható, hogy a 93% -7% -os elosztás rendelkezik a legjobb Pontosság és Felidézés értékekkel, a Felidézés esetén viszonylag nagyobb különbséggel. Napjainkban a nemzetközi életben a 80% - 20% arány a legmegszokottabb, és a Felidézés esetén itt beszélhetünk a második legjobb eredményről és Pontosságban sem maradt el a 70% - 30%, illetve 60% - 40% értéktől, ezért és amiatt, mert nagyobb adathalmaz marad tesztelni, a 80% - 20%-as változatot is használtam a későbbiekben, de legjobb módszernek a 93% - 7% arányt választottam. A 93% - 7% arány előnye, hogy rendkívül sok adat van, amiből tanulhat a modellünk, hátránya ezáltal az, hogy papíron kevés teszt adatunk marad, de fontos hangsúlyozni, hogy a 7% amelyen a tesztelés bekövetkezik még mindig több, mint 14000 sornyi adatot és kb. 294 ezer adatpontot foglal magában, így bőven van min tesztelni.

A 93% - 7%-os arány az úgy nevezett Méretezési-törvény („Scaling-law”) által került kiszámításra, ez az arány a jellemzők számától függően változik.

A tréning és a teszt adatok arányának kiválasztása után itt meg kellett még állapítanunk azt is, hogy milyen mintavétellel történjen a tréning, illetve a teszt adatsor kiválasztása:

	Pontosság	Felidézés
93% - 7% véletlen	109,29%	107,88%
80% - 20% rétegzett	107,67%	99,97%
80% - 20% véletlen	107,08%	102,77%
93% - 7% rétegzett	105,84%	104,09%
80% - 20% lineáris	102,50%	100,13%
93% - 7% lineáris	100,00%	100,00%

8. Táblázat A megfelelő mintavételi mód kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

Ahogy a fenti táblázatban is látszik ismételt a Méretezési-törvény alapuló mintavétel hozta a legjobb eredményeket a Felidézés és a Pontosság esetében is. A táblázatból az is látszódik, hogy a véletlen mintavétel jobb Felidézést hozott a 80% - 20%-os arány esetén is mint a rétegzett változat, Pontosság esetében pedig valamivel több, mint fél százalékpontnyi különbség volt csak a 80% - 20% arány esetén, míg 93% - 7%-os változat esetén érezhető különbség volt tapasztalható. A véletlen mintavétel előnye, hogy különböző lefuttatások esetén így tesztelhetünk a legkülönbözőbb adatsorokra, a rétegzett mintavétel esetén nagyobb eséllyel kerül lefuttatásra ugyanaz az adatsor. A véletlen mintavétel hátránya, hogy adott esetben félre tanulhat a modell, ha például nagyon szélsőséges adatokat kap meg (hiszen a kiugró értékek kezelése után is maradnak alacsony és magas értékek is), azonban, ahogy az adatokból láthatjuk a nagy mennyiségű adat miatt alapvetően ez nem jelent meg problémaként, kisebb adatbázisok esetén azonban ez mindenképpen problémás lehet.

IV.7. Alulmintázás

A következő lépés az volt, hogy megoldjam azt a problémát, hogy kiegyensúlyozatlan az adatbázis. Akkor beszélhetünk kiegyensúlyozatlan adatbázisról, ha egy jellemzőn belül van olyan vagy vannak olyan felvehető értékek, amelyek a másikon vagy másikonál lényegesen többször szerepelnek. Erre többféle megoldás is létezik (felülmintázás, alulmintázás, SMOTE, különböző büntetések alkalmazása stb), azonban megfelelően nagy adatbázis esetén elsőként mindig érdemes az alulmintázással (azaz undersamplinggel) élni, hiszen ezáltal adatokat veszünk, de nem veszélyeztetjük azt, hogy túl kevés adat maradjon a modellünkben, amiből tanulhat. Nagyon fontos, hogy csak a már elválasztott tréning adatsoron szabad változtatásokat eszközölni, a tréning adatsoron nem szabad változtatásokat eszközölni az adattisztítás után, ezzel elkerülve azt, hogy a modell túl vagy rosszul tanuljon.

Az alábbi táblázat szemlélteti azt, hogy különböző alulmintázási arányok esetén milyen eredményeket kaptam:

	Pontosság	Felidézés
Nincs alulmintázás (11,77% migrált)	639,67%	39,96%
18% migrált – 82% nem migrált	575,30%	41,86%
50% migrált – 50 % nem migrált	564,52%	80,86%
33% migrált – 67% nem migrált	100,00%	100,00%

9. Táblázat A megfelelő alulmintázási arány kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

Ahogy a táblázaton látható, a Felidézés, de főleg a Pontosság esetén nagyon nagy eltérések tapasztalhatók. A 33%-os aránnyal alapvető probléma, hogy nagyon rossz Pontosságot produkál minden máséhoz képest (ennek oka valószínűleg az, hogy itt már kellően nagy arányban változtak az arányok, de a kiegyenlítődésem sem következett be, emiatt „biztosra” ment a modell és kevés fogyasztót tartott felhívásra érdemesnek, így viszont sok amúgy migrálásra hajló ügyfél sem került felhívásra). A legjobb precizitást a változatlan adatsor adta, viszont itt a Felidézés volt lényegesen alacsonyabb, hiszen bár a modell a többi változathoz képest jól megtalálta, hogy kik váltanak magasabb csomagra, de sok mást is felhívott volna. Az 50%-os azaz egyenlő méretű minta („equal-size sampling”) esetén azonban bár a Pontosság valamelyest rosszabb lett, a Felidézés nagy mértékben javult, emiatt és azon kutatói megfontolásomból, hogy a kiegyensúlyozatlan adatbázist érdemes egyensúlyba hozni egy helyes, való életben jól működő modell felállításához, az egyenlő méretű mintás változatot választottam.

IV.8. Véletlen erdő

Miután a tréning és a teszt adatunk is a megfelelőképpen meg lett tisztítva, a segítségükkel elindíthattuk magát a modellt. A véletlen erdő módszer bemutatásánál említettem, hogy több különböző minőség mérési módszer („quality measure”) is létezik, ezek mellett minden fa maximális mélységét és minimális csomópontjainak számát és a kialakítandó erdő nagyságát is optimalizáltam.

	Pontosság	Felidézés
93% - 7% - gini index	106,20%	61,39%
80% - 20% - gini index	104,05%	58,48%
80% - 20% - információnyereség	102,80%	63,04%
93% - 7% - információnyereség	102,18%	64,06%
80% - 20% - nyereségarány	100,93%	96,93%
93% - 7% - nyereségarány	100,00%	100,00%

10. Táblázat A megfelelő mérési módszer kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

A fenti táblázatban háromfajta módszer került tesztelésre. Az első a közgazdaságtanban is jól ismert gini index, a második az információnyereség (information gain), a harmadik pedig a nyereségarány (gain ratio). Ezek pontos működésére jelenleg nem térek ki bonyolultságuk miatt. Az információnyereségen alapuló vágás olyan jellemzőket választ magas vágási kritériumnak, amelyek sok különböző értéket tartalmaznak, a nyereségarány pedig előbbinek a továbbfejlesztett változata. A nyereségaránnyal való számítás során a modell figyelembe veszi, hogy adott vágási kritérium választása esetén hány „ág” képződne.²⁰

Ahogy a fenti táblázaton jól látszódik, amíg a Pontosság esetén kisebb különbségek voltak, addig a Felidézés esetében már jóval több, mint érezhető, sokkal inkább óriási különbségek voltak tapasztalhatóak. Habár a nyereségarány rosszabb Pontosság értékeket hozott, mint a társai, a Felidézés esetén sokkal nagyobb százalékban képes volt kiszűrni azokat, akik valójában nem váltanának magasabb csomagra, emiatt a nyereségarányt választottam, mint mérési módszer.

A következő paraméter a modell csomópontjaira vonatkozott:

	Pontosság	Felidézés
8 db	105,49%	99,57%
16 db	105,11%	98,94%
32 db	103,44%	97,53%
4 db	101,91%	100,19%
Végtelen	101,74%	100,21%
2 db	100,00%	100,00%

11. Táblázat A csomópontok minimális számának kiválasztását segítő táblázat.

Forrás: Saját szerkesztés.

²⁰ <https://community.rapidminer.com/discussion/30340/information-gain-vs-gain-ratio>

A Pontosság értéke a csomópontok minimális számának növelésével együtt növekedett, a Felidézés pedig többé kevésbé csökkent. Megfigyelhető, hogy a 8-as érték elérése után mindkettő csökkenésnek indul, ezen észrevételre alapozva döntöttem úgy, hogy a minimális csomópontok számát 8-ban fogom meghatározni.

Ezután következett az egyes fák maximális mélységének a meghatározása. A fa maximális mélysége alatt azt értjük, hogy a gyökér csomópontjától a legtávolabb lévő csomópont hány éltre (hány szinttel lejjebb) helyezkedik el tőle²¹:

	Pontosság	Felidézés
Végtelen	1220,16%	97,23%
32 db	1204,44%	97,75%
64 db	1199,60%	98,04%
16 db	1124,60%	97,59%
8 db	100,00%	100,00%

12. Táblázat A döntési fák maximális mélységének meghatározását segítő táblázat.

Forrás: Saját szerkesztés.

Ez a táblázat már sokkal érdekesebb, hiszen megfigyelhetjük, hogy amennyiben maximális mélységet kívánunk meghatározni, ezzel óatosan kell bánni, hiszen amennyiben túl alacsony értékeket határozzunk meg, akkor - habár egy jó Felidézés értéket kaphatunk- gyakorlatilag a pontosság teljesen elromlik és szinte senkit nem fog felfelé migrálónak definiálni a modell. Ezen észrevétel és a fenti táblázat alapján döntöttem úgy, hogy nem vezetek be a maximális mélységre vonatkozó korlátozásokat, ezáltal minden fa képes lesz maga meghatározni, hogy milyen mélységig kíván lemenni.

A véletlen erdő utolsó beállítandó értéke az erdő mérete volt:

	Pontosság	Felidézés
128 db döntési fa	102,81%	101,11%
64 db döntési fa	101,98%	100,41%
256 db döntési fa	101,12%	100,14%
32 db döntési fa	100,00%	100,00%

13. Táblázat A véletlen erdő méretének meghatározását segítő táblázat.

Forrás: Saját szerkesztés.

Ahogy a szakirodalmi elemzésem során kiderült, 64 fa után már alapvetően nem voltak nagy eltérések az erdők eredményei között, 128 fás erdők felett pedig teljesen elhanyagolható különbségek keletkeztek, azonban a lefutási idő lineárisan növekedett. A táblázatból jól

²¹ <https://stackoverflow.com/questions/2603692/what-is-the-difference-between-tree-depth-and-height>

látszódik, hogy míg a 32 fás erdőről a 64-esre majd pedig arról a 128-asra lépve valamekkora növekedést mutat, addig a 256 fás erdő esetén már visszaesés tapasztalható. Kiemelném, hogy ez 4 lefuttatás átlaga és amennyiben lefuttatnánk 100 vagy 1000 modellt, valószínűleg elhanyagolható különbségeket tapasztalhatnánk. Mindezek alapján a 128 fás erdőt választottam, mivel itt a lefuttatási idő sem volt kimagasló és jó eredmények voltak tapasztalhatók.

V. Kutatási eredmények

V.1. A modell Pontosságának (Precision) és Felidézésének (Recall) kiértékelése

A következőkben a fent felállított modell eredményeit mutatom be és értékelem azokat. Először a legegyszerűbb (de ezáltal a legfontosabb) eredményeket mutatom be, utána pedig a legfontosabb és legkevésbé fontos jellemzőket. Ezután egy eset segítségével bemutatom, milyen pénzügyi és nem pénzügyi eredmények várhatóak, mik a legnagyobb kockázatok és, hogyan fejleszthető tovább a modell.

Nézzük tehát a fenti beállítások tökéletesítés után létrejövő modell eredményeit:

	Pontosság	Felidézés
#0 lefuttatás	81,10%	90,80%
#1 lefuttatás	79,90%	89,30%
#2 lefuttatás	81,40%	91,30%
#3 lefuttatás	81,40%	90,50%
#4 lefuttatás	81,50%	91,00%
#5 lefuttatás	80,30%	90,40%
#6 lefuttatás	82,10%	92,40%
#7 lefuttatás	81,30%	89,20%
#8 lefuttatás	82,10%	89,30%
#9 lefuttatás	82,40%	90,20%
Átlag:	81,40%	90,40%
Minimum:	79,90%	89,20%
Maximum:	82,40%	92,40%

14. Táblázat A végső modell lefutási eredményeit összefoglaló táblázat.

Forrás: Saját szerkesztés.

Ahogy a fenti táblázaton látható, a modell kimagaslóan jó eredményeket ért el mind a 10 lefuttatás során. Amennyiben a minimum értékeket nézzük akkor is képes volt 79% fölötti pontossággal megtalálni azokat, akik ténylegesen hajlandóak voltak váltani magasabb besorolású internetcsomagra. Ez a projekt esetében egy jó, de nem kimagasló értéknek

mondható. Azonban a Felidézés érték már sokkal kimagaslóbb, hiszen szintén minimum értéket nézve 89,2%-os eredményt ért el a modell a teszt adatsoron azaz az összes általa felhívásra javasolt ügyfél közül 89,2%-uk valójában migrált volna fölfelé. Kontextusba helyezve fontos kiemelni, hogy jelenleg a heurisztikus call centeres dolgozók sikerességi rátája az internetes felfelé migrációt tekintve kb. 26,16% volt az összes az adatbázisban regisztrált hívásra vonatkozóan.

Nem csak a rossz, hanem a jó eredmények esetén is kritikusnak kell lenni, így tettem én is. A modellt ellenőriztem hibák, vagy bármilyen olyan módosítás után kutatva, amely elferdíthetné az eredményt. Ilyenek lehetnek például a hiányzó értékek manipulatív kezelése, a teszt és tréning adatok helytelen elválasztása vagy a jellemzők nem megfelelő szűrése. Ilyen azonban a modellben többszöri ellenőrzés után sem történt, emellett a tízszeres lefuttatás is azt hivatott mutatni, hogy a különböző lefuttatások(iterációk) között nem történnek óriási különbségek, mint például kiugróan magas majd kiugróan alacsony eredmények.

Mit várhatunk ezekután a valóságban? Képes a modell tényleg megtalálni több, mint 79%-át a migrálni akaró ügyfeleknek, akár olyanokat is, akik még ők maguk sem tudják, hogy magasabb csomagra váltanának? Tényleg 100 hívásból 89 sikeres lesz? Ezen kérdésekre akkor adhatunk pontos választ, ha bevezetjük a való életben a modellt, majd egy hónap, negyedév vagy éppen egy év után összehasonlítjuk a sikerességi rátát az elmúlt években tapasztalhatókkal. Azonban lényeges kiemelni, hogy a teszt adaton a modell pontosan azt csinálta, mint amit nem laboratóriumi körülmények között, újonnan az adatbázishoz rendelt adatokkal tenne. Számára teljesen ismeretlen adatsorokat kapott, olyanokat, ahol nem tudta, hogy mi a végkimenetel. Teljes mértékben az általa rendelkezésre álló jellemzők alapján becsült, és ezek között ugyanúgy fordulhattak elő hiányzó értékek, mint ahogy a való életben is- hiszen ezek szintén való életből vett, valós hívások adatai voltak, „csak” a végkimenetelt tartottam vissza a modelltől egészen az értékelésig.

A nemzetközi gyakorlatban is több, hasonló modellt találhatunk, amelyek az én általam fejlesztett modellnél még sokkal pontosabb predikciót is képesek alkotni például a szintén összetett HR területén.²² Ezek alapján a precizitáshoz kötődő mérési eredmények egyáltalán nem számítanak olyan meglepőnek, mint elsőnek tűnhetnek. A Felidézés esetében a 89% fölötti érték rendkívül kimagasló eredménynek számít a call center alapú kereszt és felülértékesítés

²²<https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html>

területén. (Összehasonlításuképpen a mesterséges intelligencia nélkül az internet migráció tekintetében [ahol ráadásul legmagasabb volt ez a százalékos érték] sem érte el a 30%-ot a hívások sikeressége.) Emiatt úgy ítélt meg, hogy fontos felkészülni azokra az esetekre, amikor, valamilyen oknál fogva rosszabbul működne a modell a gyakorlatban, ezért három forgatókönyvet alkottam meg. A vélt legjobb kimenetel (best case scenario) lesz a 89%-os értékelés, emellett viszont meg fogom vizsgálni az eredményeket 70 és 55% esetén is.

Ezek pusztán biztonsági esetek, a modell létrehozása során összesen kb. 400-500 lefuttatást végeztem, törekedve a lehető legpontosabb és legjobb modell kialakítására. Tekintsünk tehát ezen százalékokra inkább úgy, mikor egy üzleti tervezés után tervbe vesszük azt, hogy egy projekt költségei X százalékkal túlmutatnak azon, amelyekkel előre kalkuláltunk és ez milyen hatással lehet a különböző mutatóinkra.

Mielőtt ezen eredményeket egy fiktív esettanulmány esetében szemléltetném azelőtt, egy 128 fás véletlen erdőn keresztül bemutatom azon jellemzőket, amelyek a legfontosabbnak minősültek a fák szerint, illetve kiemelek olyanokat, amelyek adott esetben befolyásolhatónak tűnnek, azonban a gyakorlatban még sem váltak meghatározóvá.

V.2. Legjelentősebb vágó kritériumok

Összesen 182 jellemzővel rendelkezett adatbázisunk, a kiszűrési folyamat után ebből 166 jellemző maradt, amelyekből ténylegesen tanulhatott a véletlen erdő.

A jellemzők fontosságának megállapításához megvizsgáltam, hogy a jellemzők hányszor kerültek kiválasztásra magas szintű vágási pontként. Magas szintű vágási pontnak azt nevezzük, amikor az adott jellemző vagy a gyökér csomópontban vagy az alatta lévő két szinten helyezkedik el, azaz az adott fa által nagy mértékben képes eldönteni, hogy mely fogyasztók váltottak magasabb internetcsomagra és melyek nem.

A 166 jellemző közül 90 olyan jellemző volt, amely a 128 fából egyetlen esetben sem volt az első három vágó kritérium között, 18 nem került egy esetben sem az első kettő közé, 21 volt maximum második szintű vágó kritérium és végül a 166 jellemzőből 37 volt olyan, amelyet legalább egy döntési fa választott a legmagasabb szintű vágó kritériumból. Ebből jól látszik, hogy mennyire változatos módon gondolkodtak a fák, mennyire összetett és komplex a probléma (egy-egy egyedi döntési fa mennyire nem adhat pontos eredményt) illetve azt is, hogy hiába volt több, mint 160 jellemzőnk, a modell sikeresen tudott szelektálni ezek közül.

Jellemző név	Első szintű vágási pontnak választások száma:	Második szintű vágási pontnak választások száma:	Harmadik szintű vágási pontnak választások száma:
Internet csomag ára	13	10	6
Internet hűségidő	11	7	19
Kapcsolatfelvételek száma	10	6	14
Gyűjtött internetes migrációs pontok	8	18	35
Egyenleg	8	7	12
Idős férfi	7	8	6
Elérhető televíziós migrációs pontok	7	6	3
Televízió hűségidő	7	5	7
Átlagos hívás hossz	6	8	13
Utolsó tranzakció ideje	5	10	16

15. Táblázat A véletlen erdő által legfontosabbnak talált jellemzők.

Forrás: Saját szerkesztés.

A fenti táblázatban a véletlen erdő által legjobb vélt jellemzők láthatók, illetve az, hogy hányszor lettek választva első, második vagy harmadik számú vágó pontnak. A következőben ezen jellemzőket értelmezem.

A véletlen erdő modell által legtöbbször első számú vágó kritériumnak választott jellemző az „Internet csomag ára” volt, amely az adott időszakban az adott előfizető élő otthoni internet termékeinek az értékét jelöli forintban kifejezve. Választ kaphat a modell a jellemző segítségével arra, hogy mennyire érzékenyek egyes fogyasztóink az árak változására, amely szintén könnyítheti a behatárolását a migráló ügyfeleknek. Ezen jellemző segítségével a modell képes lehet meghatározni azokat a fogyasztókat, akik már most is a legmagasabb csomagban vannak, azaz azokat, akik már nem is tudnak felfelé migrálni. Továbbá a más termékekre vonatkozó érték jellemzők segítségével azokat az ügyfeleket is észlelheti a modell, akik például sokat fizetnek TV és telefon szolgáltatásra, de internetre nem, azaz valószínűleg vagyonszebbek vagy a szórakozás fontos számukra, ezáltal pedig az internet felfelé migrációra is magasabb hajlandóságuk mutatkozhat.

A második legtöbbször első szintűnek választott vágó kritérium volt az „Internet hűségidő” azaz, hogy mikor jár le a hűségideje az ügyfél által rendelkezésre álló otthoni internet terméknek napban kifejezve. Amennyiben valakinek még hosszú hűsége van egy adott internetcsomagra, abban az esetben lehetséges, hogy kevésbé hajlandó váltani egy nagyobbra,

hiszen nemrég választotta valamilyen okból a jelenlegi csomagját, amelyre azért vállalat feltehetőleg hűséget, mert eleget tesz a jelenlegi elvárásainak. Azonban, akinek már a hűségideje közeledik a lejáráthoz, az könnyen lehet, hogy éppen azon gondolkodik, hogy milyen csomagba lépjen át, lehetséges, hogy a csomag, amiben van már nem is létezik és sokkal jobb konstrukciók váltak elérhetővé, amelyekről az ügyfél adott esetben nem is tud.

A következő listán szereplő jellemző a „Kapcsolatfelvételek száma”, amely azt mutatja meg, hogy az adott ügyfél hányszor lépett a vállalattal összesen kapcsolatba. Ebbe beleszámít az is, mikor az ügyfélszolgálat kereste meg, de az is, mikor ő tett hibabejelentést, érdeklődött valamilyen termékek iránt, panaszt tett vagy visszajelzést küldött. Ezen jellemző elől szereplése alapján feltételezhetjük például, hogy amennyiben egy fogyasztó sokat lép kapcsolatba a vállalatunkkal, akkor máshogy fog viselkedni az internet migráció tekintetében, mint azok, akik kevésbé. Ennek oka lehet például az, hogy sokszor kapcsolatba lépők szívesebben váltanak nagyobb internetkapcsolatra, amiatt, mert a gyakori kapcsolattartás végett életük szerves részét képezi a vállalat és szeretnének, gyorsabb, magasabb minőségű szolgáltatásokat.

A következő a „Gyűjtött internetes migrációs pontok” jellemző azt jelöli, hogy az adott ügyfél eddig összesen hány migrációs pontot ért el az otthoni internet előfizetéséhez kötődő felfele migrációval. Migrációs pont akkor keletkezik, ha adott ügyfél egy csomagról egy annál magasabb csomagra vált: minél nagyobb csomagra vált, annál több migrációs pontja keletkezik, és ezen migrációkkal gyűjtött pontokat különböző promóciós termékekre válthatja be. A jellemző esetében alapvetően a migrálási hajlandóságot állapíthatjuk meg, hiszen akinek magas az itt szereplő pontszáma az azt jelenti, hogy korábban többször és / vagy nagy mértékben cserélte magasabb kategóriájúra az internet csomagját. Ezen jellemző esetében kiemelném, hogy rendkívül sokszor lett második és harmadik szintű vágási kritérium, amelyből az látszik, hogy valamilyen jellemző vagy jellemzők mentén történő vágás után kifejezetten nagy mértékben dönthető el ezen egy jellemző alapján, hogy kik hajlandóbbak migrálni és kik nem.

Az „Egyenleg” jellemző az adott ügyfél adott napon érvényes egyenlegét mutatja meg forintban, mely érték felvehet pozitív és negatív értéket is. Itt példaként előfordulhat az, hogy a tartozással rendelkező ügyfelek kevesebb eséllyel hajlandóak felfelé migrálni, mint a tartozással nem rendelkezők.

Az „Idős férfi” egy nagyon érdekes, utólagosan generált jellemző. A jellemző, mint egy „flag”, azaz zászlóként működik, tehát „1”-essel jelöli azokat az ügyfeleket, akik 69 évnél idősebbek és férfiak (akik nem azok esetében „0” értéket vesz fel a jellemző). A jellemző

jelenlétének az oka az lehet, hogy a call centerben dolgozók már korábban megfigyelték azt, hogy az idősebb férfiak valamelyest homogén módon viselkednek és ezen jellemző bevonásával sokkal könnyebben és gyorsabban képesek beazonosítani, hogy mikor telefonál egy idős férfi vagy terveznek felhívni őket egy idős férfit éppen. Esetükben lehetséges, hogy számukra a legalapvetőbb (vagy éppen semmilyen) internet szolgáltatás is elégséges (vagy adott esetben csak az megfizethető), más részről pedig az is elképzelhető, hogy könnyen rábeszélhetőek a call centeres értékesítők által a váltásra.

A listán következő két jellemző megjelenése elsőre meglepetést válthat ki, hiszen az internetre vonatkozó „társaikkal” már találkoztunk fentebb. Az első jellemző az „Elérhető televíziós migrációs pontok”, azaz az a jellemző, amely megmutatja, hogy a fentebb bemutatott migrációs pontok közül az adott ügyfél a televíziós termékre vonatkozó felfelé migrációval összesen hány pontot érhet el (tehát a „Gyűjtött internetes migrációs pontok” jellemzővel ellentétben nem múlt, hanem sokkal inkább jövőcentrikus jellemzőről van szó). Ezen jellemző jelenlétére magyarázat lehet egy kettős felülrértékesítés jelenléte, miszerint, amennyiben nagy migrációra van lehetőség a televízió területén és az internet területén is, akkor egy csomagajánlat érdekes lehet az ügyfél számára, vagy éppen előtérbe kerülhet az is, hogy, akinek alacsony a maximálisan elérhető migrációs pontja a televízió előfizetés esetén, annak drága csomagja van és adott esetben emiatt hajlandó az internet csomagját is magasra váltani. Tehát itt is megjelenik a multimédiás szolgáltatásokra való költési hajlandóság, illetve szerepet kap az anyagi helyzet is, amellyel az „Internet csomag ára” jellemző „segítségére” lehet ez a jellemző, mint egy alacsonyabb szintű vágási kritérium, ahogy azon jellemző esetén kifejtettem.

A másik televízió előfizetéshez kapcsolódó jellemző a „Televízió hűségidő”, amely az „Internet hűségidő”-höz hasonlóan azt mutatja meg, hogy az adott ügyfél TV előfizetéséhez tartozó hűségéből még hány nap van hátra. Ezen jellemző jelenlétét magyarázhatja az, hogy adott esetben egy-egy előfizetés lejártá után vált inkább vagy egyáltalán hallgat meg jobban ajánlatokat egy-egy fogyasztó más fajta termékekre vonatkozóan is, akkor érthetővé válik, hogy miért kerül ilyen előre ez a jellemző. Lehetséges, hogy az is szerepet játszik ebben, hogy manapság egy fiatal felhasználó akár a TV előfizetését fel is függeszti, de ellenben egy magasabb internetcsomagra vált, hogy jó minőségben tudjon különböző streaming szolgáltatásokat igénybe venni.

A televízió előfizetéshez kapcsolódókon kívül további két olyan jellemző került az első tízbe, amelyek esetében a modellen kívül nehéz megállapítani, hogy pontosan miért tölt be ilyen

magas helyezést. Ilyen a „Utolsó tranzakció napja” azaz, hogy az utolsó tranzakció hány napja történt és az „Átlagos hívás hossz” azaz az, hogy egy adott call centerhez bejövő / kimenő hívás esetén az adott ügyfél esetén milyen hosszú volt egy átlagos hívás (másodpercben megadva). Itt előbbi esetében vonhatunk gazdaságpszichológiai párhuzamokat, mint például, hogy amennyiben éppen egy számlát fizetett be pár napon belül az ügyfél, akkor nem feltétlen akar még több pénzt költeni a vállalat felé, esetenként előfordulhat az is, hogy egy-egy hosszabb beszélgetés során bonyolultabb panaszok vagy hívások jelentkeznek, lassabb a problémamegoldás, amely miatt kisebb valószínűséggel fog migrálni a fogyasztó.

A véletlen erdő és más gépi tanuló algoritmusok előnye, hogy felismernek ilyen rejtett összefüggéseket is (mint például az előbb említett négy jellemző esetében), amelyek más körülmények között, akár alapos kutatások során sem derülnének ki, mint fontos tényezők egy felülértékesítési probléma esetén.

V.3. Legkevésbé jelentősebb vágó kritériumok

A legkisebb arányban magas szintű vágókritériumnak választott jellemzők közül a számomra meglepőbbeket három csoportba soroltam, melyeket röviden az előbb említett csoportokra bontva mutatok be:

Az első ilyen csoport a kapcsolattartási jellemzők csoportja. Eredendően könnyen feltételezhetnénk azt, hogy a migrálási hajlandóság összefüggésben van azzal, hogy az adott ügyfél hányszor keresett minket, érdeklődött bizonyos termékekért, vagy éppen jelentett be hibákat. Ez érthetőnek és igazoltnak is tekinthető, mivel a harmadik legerősebb vágó kritérium volt az, hogy az adott ügyfél hányszor került kapcsolatba velünk, azonban ha ezt szétdaraboljuk, attól lényeges eltéréseket nem tapasztalhatók a modellben. Itt megjelenik többek között a: bejelentett panaszok száma, termékek iránti érdeklődések száma, a hibabejelentések száma. Továbbá azok a jellemzők, amelyek meghatározzák azt, hogy pontosan hány nappal ezelőtt történt a legutóbbi hibabejelentés, panasz, termék iránti érdeklődés vagy adott szolgáltatáshoz tartozó visszajelző levél beérkezése.

Mindez fontos kérdéseket vethet fel, hiszen kiemelten a szolgáltatások tekintetében rendkívül nagy figyelmet fordítunk a panaszkezelésre, azonban jelen modell alapján vagy a vizsgált vállalat panaszkezelési módszerei nagyon közömbösek (se nem nagyon rossz, se nem nagyon pozitív) vagy hiába panaszkodnak az ügyfelek, akkor is hajlandóak felfelé váltani. Ezt tehetik akár abban a hitben, hogy a magasabb szintű internetkapcsolattal kevesebb problémájuk lesz (és az elégedett ügyfelek esetében pedig pont hiába elégedettek adott esetben, azért nem

migrálnak felfelé), ezekkel szemben ezek ellentétjei (amelyeket nagyobb valószínűséggel várnánk) is előfordulnak, pontosan emiatt nem válnak kiemelten fontossá a modell számára a fent említett kapcsolattartáshoz kötődő jellemzők. Ebben az esetben (amennyiben ezeken a területeken kívánunk javításokat végezni) érdemes lehet mélyebre ásni valamilyen kvalitatív (mélyinterjú, fókuszcsoportos) módszerrel annak érdekében, hogy jobban megtudhassuk, mi állhat ennek a hátterében.

Szintén gyengén befolyásoló jellemzők az EDM-re azaz az elektronikus direkt marketingre vonatkozó link kattintási arány. Az talán egyértelmű, hogy nem lesz hatással a migrációra az, hogy a fogyasztó milyen eszközön nyitja meg a hírleveleinket. Azonban az, hogy rákattintott-e a benne szereplő linkre, vagy nem az már sokkal kevésbé egyértelmű. Azért is érdekes ez, mert azok a jellemzők, amelyek azt határozták meg, hogy az adott ügyfél kapott-e EDM-et vagy megnyitotta-e azt, szerepeltek bizonyos fák legfelsőbb 3 vágókritériumai között. Tehát azok esetében a modell nagyobb jelentőséggel képes volt az adott jellemzők alapján eldönteni, hogy ki fog az internet termékére vonatkozóan felfelé migrálni és ki nem.

A harmadik érdekes megállapítás azon jellemzőkhöz köthető, amelyek a valamilyen más szolgáltató általi veszélyeztetettséget vizsgálnak. Itt három kiszolgáló volt konkrét jellemzőkkel mérve, olyanok, akik mind rendelkeznek otthoni internetszolgáltatással. Feltételezhetnénk, hogy a kompetitívabb területek esetében a migrálási hajlandóság alacsonyabb lehet (hiszen több opciójuk van, átmehetnek másik szolgáltatóhoz), de a döntési fák és a véletlen erdő nem sorolta előtérbe ezeket. Ebben ezek mellett szerepet játszhat az is, hogy amennyiben veszélyeztetett területen lakik egy ügyfelünk, az nem kerül mérésre, hogy adott szolgáltatások esetében a versenytárs milyen árakkal dolgozik. Amennyiben ezt mérnénk, könnyen lehet, hogy abban az esetben a modell már ezek segítségével (vagy egy ezekből általunk kreált harmadik jellemző) képes lenne még jobb előrejelzést adni és ezen adatokból is értékes információkat kinyerni.

V.4. Eredmények kiértékelése

Mint látható az általam épített véletlen erdő egészen széles skálán gondolkodott és nagyon lényegesnek tartom kiemelni, hogy ezek csak a legfontosabb és a legkevésbé fontosabb jellemzők bemutatása volt. A véletlen erdőnek alapvetően nagy előnye, hogy rengeteg jellemzőt tud egyszerre kezelni, sorba állítani őket és ez alapján gondolkodni. Teszi mindezt úgy, hogy a képzési folyamata végén konkrét eredményeket közöl azzal kapcsolatban, hogy melyek azok az ügyfelek, akiket érdemes megcélozni és kik azok, akiket nem. Ezután nincs szükség

tüzetesebb elemzésre, hogy melyik jellemzők és esetlegesen miért kerültek kiválasztásra, hanem akár azonnal és automatizáltan jelezhető a call center felé, hogy bizonyos termékekkel kapcsolatban kiket érdemes megcélozni. Ahogy elhangzott a modell összességében az otthoni internet előfizetésnek a felfele migrációjára irányult, de néhány plusz jellemző megadásával konkretizálható az is, hogy adott fogyasztónak pontosan melyik termékeket érdemes ajánlani, adott esetben kimutatva ezen termékek felajánlási sorrendjét is (ehhez mindösszesen az alapján kell sorba állítani a termékeket, hogy mennyire magabiztos a modell abban, hogy az adott terméket megvásárolná a fogyasztó).

Rengeteg időt, energiát, humán és pénzügyi erőforrást, számítási kapacitást tud egy ilyen modell megspórolni, hiszen míg más módszerek alkalmazása során adott esetben csak egy-egy jellemzővel kapcsolatban tudjuk meg, hogy mennyire lényegesek, utána kell járni annak, hogy egyes fogyasztók esetén pontosan hogyan is jelennek meg, esetlegesen képletekbe kell behelyettesíteni vagy más kutatásokat is be kell vonni. ahogy fentebb is említettem, itt is amennyiben a menedzsment szükségesnek érzi különböző kutatásokban ötvözheti ezt a módszert is, azonban ez nem szükséges, nem kell minden ügyfélnek külön utánanézni vagy akár perszónákba sorolni őket, mivel a modell ezeket mind elvégzi helyettünk. Ehhez szükség van a megfelelő adatokra, a helyes beállításokra, pontos tesztelésre, implementálásra és karbantartásra.

V.5. Megjegyzések

Az adatok tekintetében érdemes a lehető legtöbb gyűjteni és rendszerezni, addig amíg a józan ész megkívánja és ezzel nem sértünk meg törvényeket, illetve nem lesz túlságosan költséges akár a gyűjtésük, akár a tárolásuk. Egy modellnek mindig könnyebb jó adatokból tanulnia, mint rosszakból és adott esetben egy-egy jelentéktelennek tűnő adat, észrevétel is lehet értékes jellemző egy modell számára. Példaként említem egy adott vállalat Facebook oldalára beérkezett üzeneteket, ezek eredendően szövegesek, azonban rengeteg előre megírt algoritmus létezik, amelyek egy adott szótár segítségével képesek megállapítani, hogy melyek a pozitív üzenetek és melyek nem. Azonban ebben az esetben is mielőtt kidolgoznánk ezt a megoldást, érdemes megnézni, hogy egyáltalán mennyire használják az ügyfelek a csatornát, mivel, ha alig kapunk ilyen üzeneteket, akkor valószínűleg nem ezt kell a prioritás előterébe helyezni.

Amennyiben a vállalatunk még az adatgyűjtés és adott esetben a digitalizáció kezdetén áll, érdemes nemzetközi legjobb gyakorlatok után olvasni, vagy külsős konzulánsokat felbérelni, akik segítenek a vállalatnak abban, hogy adatvezérelten gondolkodhassanak. A

fentebb említett jellemzők betekintést nyújthatnak abba, hogy melyek lehetnek a legfontosabbak, de a modell folyamatosan fejlődik, minél több adatot kap. Minden munkanap egy lefuttatást igényel, hogy megtudhassuk az adott pillanatban kiket érdemes célozni, és ahogy akár a vevőkörünk változik, változhatnak a legfontosabb jellemzők is. Amennyiben azt figyeljük meg, hogy akár egyik évről a másikra a modell kevésbé megbízható, rosszabb eredményeket mutat, akkor érdemes arra gondolni, hogy lehetséges, hogy új jellemzőkre lenne szükség, amelyek jelenleg nincsenek mérve, vagy meglévő jellemzők esetén sok hiányzó érték keletkezett különböző problémák miatt, amelyeket érdemes minél előbb orvosolni és nézni, hogy az eredményesség miképpen változik. Ebben az esetben szintén jó példa a Facebook megemlézése, hiszen a Facebook indulásakor még kevés vállalat jelent meg rajta, azonban miután egyre nagyobb teret nyert magának a platform, úgy vált egyre fontosabbá az ottani adatok gyűjtése és amennyiben ez hiányzott, abban az esetben romolhatott a modell értéke. Lehetséges, hogy amivel eddig a call centert keresték vagy e-mailt írtak az ügyfelek, ma már ezt más felületen teszik, ha érdeklődnek termékek iránt vagy éppen panaszt nyújtanak be.

VI. Magyar Telekom esettanulmány

Ahogy az eredményeknél jeleztem a várható pénzügyi eredmények és üzleti hatások esetén a számolás tekintetében 3 különböző lehetséges változatot is megvizsgállok. Az első eset a modell eredeti eredményein épül, azaz 79,9%-os precizitással és 89,2%-os Felidézéssel számolunk, ebben az esetben („A” változat), a rosszabb („B”) esetben 70%-kal, az általunk feltételezett legrosszabb esetben („C”) pedig 55%-kal.

Először is fontos meghatározni azt, hogy mely számok azok, amelyek változását mérni akarjuk, majd pedig alapokat kell találni, amelyeken felmérhetjük, hogy összegében és százalékban mit jelenthetne egyes vállalatok esetében. Utóbbi esetben persze kézen fekvő lenne az adatbázist biztosító vállalat adataival számolni, de ezt biztonságtechnikai okokból nem fogom megtenni, helyette cikkek, illetve irodalmakban meghatározott arányokat, számokat fogok segítségül hívni, illetve egy konkrét vállalatot kiválasztani, hogy szemléltessem, mennyit jelenthetnek ezek a változtatások pontosan.

A következő értékek, amelyek meghatározására törekszem:

- Bevételekre gyakorolt hatás
- Profitra gyakorolt hatás
- Költségekre gyakorolt hatás

- ROMI-ra gyakorolt hatás

Konkrét vállalatnak a Magyar Telekom Nyrt-t (Telekom) választottam, mivel úgy ítélt meg, hogy az ő esetükben több a számításhoz fontos összeget sikerülhet meghatározni. A számításokhoz a Telekom 2016-os éves konszolidált jelentését vettem alapul, abból a megfontolásból, mert a 2019-es beszámolóval ellentétben, itt valamelyest kitérést mutatnak arra, hogy a kereszt és felülrértékesítésre mekkora keretet költhetnek. A vállalat marketing költségvetése 12,6 milliárd, az ügynöki jutalék 7,1 milliárd forint. A vállalat bevételei 602,7 milliárd forintot tettek ki, a vállalat működési eredménye pedig 79,6 milliárd forint volt.²³ Ezen eredmények három ország összesen 10 982 419 előfizetésére vonatkoztak.²⁴ Az előfizetők számában a közel 11 millió előfizetőn belül számunkra kiemelten fontos az összesen 1 289 806 fő, akik otthoni internet előfizetéssel rendelkeznek Magyarország, Montenegró vagy Észak-Macedónia területén.

A következő számítások alapvetően a fenti két dokumentum és a saját vállalatomnál tapasztaltak alapján kerültek kiszámításra. Ahol ezeken felül és saját gondolataim mellett extra forrásokat tartalmaz az elemzésem, ott ezt jelzem. Ezen „eset” előfeltételezése, hogy 2016-ban a telesales heurisztikus alapon működött és az ügyfélszolgálat döntötte el, hogy kit hívott, nem pedig valamilyen modell alapján határoztak erről. Azonban amennyiben ez nem így volt sem befolyásolja ennek a résznek a jogosságát, hiszen a lényege a fejezetnek, hogy a modell ereje, illetve eredményei valós számokon alapulva, érthetően kerülhessenek bemutatásra.

Az első táblázatban ismertetem azt, hogy a Telekom esetében jelenleg kb. hány otthoni internet migrációt célzó call centeres hívás történik, illetve ebből mennyi lehet a sikeres. Ezt hasonlítom össze azokkal az adatokkal, amelyek az általam felvázolt „A”, „B” és „C” forgatókönyvek által jönnek létre:

	Sikeres hívás (db/év):	Összes hívás (db/év):
Jelenlegi állapot	321 386	1 228 540
"A" forgatókönyv	256 787	287 878
"B" forgatókönyv	256 787	366 839
"C" forgatókönyv	256 787	466 886

16. Táblázat A különböző forgatókönyveket bemutató táblázat.

Forrás: Saját szerkesztés.

²³ https://www.telekom.hu/static-tr/sw/file/IFRS_Csoport_2016.pdf

²⁴ https://www.telekom.hu/static-tr/sw/file/4q_2016_hun.xlsx

A táblázat esetén két rendkívül lényeges következtetést lehet leszűrni. Az egyik megállapítás az, hogy a véletlen erdő modell alapján kidolgozott módszer esetén a sikeres hívásszám lecsökken, mivel arányaiban a modell 20%-kal kevesebb migráló ügyfelet képes megtalálni, mint a jelenlegi heurisztikus, tapasztalati alapon működő szelekció és telefonálás. ez probléma, amelynek indirekt „orvoslására” néhány bekezdés múlva ki fogok térni.

A második oszlopban látottak esetén viszont már egy sokkal szimpatikusabb eset látható, itt kirajzolódik, hogy drasztikusan lecsökkent azon hívások száma, amelyek szükségesek ahhoz, hogy megtaláljuk a migráló ügyfeleket, még az általam legrosszabbnak titulált „C” esetben is. Ezek alapján ambivalens érzéseink lehetnek, hiszen egy részről a várható bevételeink csökkennek, viszont a költségeink is csökkennek. Emiatt megvizsgáltam, hogy van-e lehetőség úgy növelni a sikeres értékesítések számát, hogy a hívásaink sikerességi aránya is nő a kiinduló állapothoz képest.

Vállalatom hívási adatainak vizsgálatakor megnéztem, hogy egy adott évben összesen az ügyfelek hány százaléka kerül felkeresésre. (Családi körökből tudom, hogy van, akit akár évekig nem, mást viszont évente többször is felhívunk.)

Ebből kiderült, hogy az ügyfelek mintegy 31%-ának felhívására nem került sor egy adott évben, történt ez azért, mert a call center tapasztalati alapon döntve inkább más ügyfeleket hívott fel többször (a mi modellünk is, ha úgy ítéli meg, felhívja többször ugyanazon vásárlókat egy adott időszakon belül). Azt kihasználva, hogy mind a három esetben az összes hívásszám esetén drasztikus csökkenés megy végbe, lehetőség nyílik arra, hogy felhívjuk a maradék 31%-át az ügyfeleknek. Itt kb. 399 839 ügyfelet jelentenek azok, akik nem kerültek felhívásra, közülük kb. 104 597 lenne az, aki magasabb internet előfizetésre váltana és a modell közülük kb. 83 573 ügyfelet tud megtalálni, ennyivel nő tehát a sikeres hívások száma. Emellett a sikertelen hívások száma is nőni fog, amelyek száma az összesítő táblázatbeli értékekhez szintén hozzáadásra került. Előtte viszont ismertem a fejlesztési költségeket és kitérek röviden a jutalékokra, illetve az egy migrációra jutó bevételre.

A fejlesztési költségek esetén alapvetően abból indultam ki, hogy számomra mennyi idő volt a modell kifejlesztése (amely nálam sokkal tapasztaltabb dolgozók esetén, minden valószínűség szerint rövidebb időt vesz igénybe vagy pedig jobb modell keletkezik), majd ehhez az időhöz hozzáadtam, hogy a különböző adattárházak összekötése, az adatok további előkészítése mennyi időt vehet igénybe, beleszámolva a dokumentációt is. Efelett további extra időket is hozzáadtam csúszás esetére.

A munkaerő költségek mellett azzal is számoltam, hogy egy ilyen nagyvállalat esetén a KNIME Analytics helyett már a KNIME Server kerülhet használatba, amely különböző extra funkciókkal rendelkezik, mint például a REST API támogatás vagy a munkafolyamatoknak a felhőben történő lefuttatásának a lehetősége. Munkaidők esetén szerződéses vállalkozói bérekkel számoltam – óraszámban megadva. A licenz esetén pedig az éves díjat vettem figyelembe. A KNIME költségei a KNIME oldaláról²⁵, a bérköltségek pedig a HAYS Hungary 2021-es bérfelméréséből származnak.²⁶

Név:	Óraszám	Órabér	Összesen
Üzleti elemző	200	12500 HUF	2 500 000 HUF
Adattudós	200	13500 HUF	2 700 000 HUF
Adatmérnök	200	14000 HUF	2 800 000 HUF
KNIME Server Medium	x	x	9 028 750 HUF
Összesen	600	-	17 028 750 HUF

17. Táblázat A projekt költségeit (a jutalék költségeken kívül)
összesítő táblázat. Forrás: Saját szerkesztés.

A humán erőforrás tekintetében tehát összesen három dolgozóval számoltam, fejenként 200 munkaórával és kb. 13-14 ezer forintos órabérrel. A fejlesztés során számoltam egy adattudóssal, aki alapvetően az adattisztítással, a modell felépítésével és értékelésével foglalkozik; egy adatmérnökkel, aki az adattárházakat térképezi fel, köti össze a folyamat elején, hogy elkezdődhessen az adattudós munkája majd pedig a végén is a már létező modellel, hogy oda jusson a kellő eredmény ahová kell. Az üzleti elemző pedig végzi a dokumentációt, garantálja, hogy a menedzsment elvárásai valósuljanak meg illetve „fordít” az IT és a marketing között.

Az összesítő táblázat esetén már csak az értékesítéshez köthető jutalékot és bevételt szükséges meghatározni az egyszerűsített modell esetében. Itt leszögezném, hogy a Telekom esetében az értékesítési ügyfélszolgálat a honlapjukon elérhető információk alapján teljesen kiszervezett²⁷, ez pedig az én vállalati tapasztalatom alapján azt jelenti, hogy az ügynöki jutalék (tehát a 7,2 milliárd) az egész telesales üzletág költségeit jelenti, nem csak a jutalékokét.

Az egy migrációra jutó bevétel esetén vizsgáltam a Telekom jelenlegi internetcsomagjait (mivel jelenleg ezekre és nem a 2016-os adatokra létezik megjelenés), amely alapján az egy migrációra jutó profitot 1493,33 forintban határoztam meg. A meghatározáshoz a különböző mértékű migrációkkal történő bevételhez rendelték elképzelhető valószínűségi

²⁵ <https://www.knime.com/knime-software/knime-server-pricing>

²⁶ https://cloud.email.hays.com/hu_salary_guide_english

²⁷ <https://www.telekom.hu/rolunk/karrier/partneri-hirdetesek/partneri-hirdetesek-20200708-1>

átlagokat és ez alapján jött ki az adott összeg. Az egyszerű számítás módszerét az alábbi táblázat mutatja:

Megnevezés	Érték	Valószínűség	Súlyozott érték
1>4	3460 HUF	5%	173 HUF
1>3	2960 HUF	10%	296 HUF
2>4	2420 HUF	15%	363 HUF
1>2	1040 HUF	25%	260 HUF
2>3	1920 HUF	20%	384 HUF
3>4	500 HUF	25%	125 HUF
Összesen		100%	1601 HUF

18. Táblázat Az egy értékesítési migrációval képződő bevétel meghatározását segítő táblázat.

Forrás: Saját szerkesztés.

A táblázatból láthatjuk, hogy egy átlagos internetes migráció 1601 forint havi bevételt hoz magával, azaz éves szinten 19 212 forintot, amelyből egy hónapnyi (a saját vállalatomnál tapasztaltak alapján) összeget tesz ki az ügynöki jutalék.

Mindezen számítások tudatában az alábbi két táblázat mutatja az esetünkben a különböző forgatókönyvek következtében bekövetkező változásokat:

Forgatókönyv	Összes hívás szám:	Sikeres hívások :	Hívás szám változás:	Hívás szám változás %:	Sikeres hívás szám változás:	Sikeres hívás szám változás (%):
"A"	381 569	340 360	- 846 971	-68,94%	18 974	5,9%
"B"	486 229	340 360	- 742 311	-60,42%	18 974	5,9%
"C"	618 836	340 360	- 609 704	-49,63%	18 974	5,9%
Jelenleg	1 228 540	321 386	-	0	-	0,0%

19. Táblázat A vállalat hívás számait összefoglaló táblázat.

Forrás: Saját szerkesztés.

Forgatókönyv	Bevétel növekedés (HUF):	Költség csökkenés (HUF):	Költség növekedés (jutalék) (HUF):	Költség növekedés (fejlesztés) (HUF):	Profit változás (HUF):
"A"	364 528 488	319 877 957	30 377 374	17 028 750	637 000 321
"B"	364 528 488	280 350 716	30 377 374	17 028 750	597 473 080
"C"	364 528 488	230 268 652	30 377 374	17 028 750	547 391 016
Jelenleg	0	0	0	0	0

20. Táblázat A vállalat pénzügyi eredményeit összefoglaló táblázat.

Forrás: Saját szerkesztés.

A táblázatokból jól látszódik, hogy azzal, hogy a véletlen erdő modell lehetőséget adott arra, hogy olyan ügyfeleket is felhívjunk, akikre eddig nem volt lehetőség, azzal az otthoni internetkapcsolathoz köthető értékesítés 5,9%-kal növekedett, amely bevételi szinten mintegy

364,5 milliós növekedést jelent éves szinten. Azonban nem csak a bevételek növekedése, hanem a költségek csökkenése is kimagasló, hiszen a legpesszimistább esetben is 230 millió forinttal csökkennek a telesales költségek. A költségcsökkentés esetén az összeg úgy került kiszámításra, hogy 7,2 milliárdos jutalék költségnek néztem meg az én vállalatomban tapasztalt arányok alapján, hogy mekkora részét teheti ki az internet migrációhoz kötődő hívások száma, és ennek eredményeképpen jött ki egy 463, 986 milliós összeg, amelyet a modell segítségével sikerült 230 millió forinttal lecsökkenteni. Ezt a csökkentést pedig azzal sikerült elérni, hogy a hívások számát a legpesszimistább esetben is 49,63%-kal csökkenthetjük a kiinduló állapothoz képest, így ennyivel kevesebb szerződött partnerre van szükségünk ehhez a telesales folyamathoz (ettől függetlenül nem feltétlen a bizonyos partnerekkel történő szerződésbontás a legjobb lépés, erre későbbiekben kitérek még).

Összességében tehát a Telekom esetében a modell segítségével, csak az internet migrációra történő alkalmazása során a legpesszimistább esetben is több, mint 547 millió forinttal nő az éves profit, amely a vállalat 79,6 milliárdos működési eredményéhez képest kb. 0,68%-os növekedést mutat. Ez ránézésre nem feltétlen számít soknak, de fontos hangsúlyozni, hogy ez mindösszesen egy termék esetén és csak a felfelé migrációra vonatkozott. Ugyanúgy véletlen erdő modell építhető a többi termékcsoporthoz, adott esetben termékszinten, amelyek segítségével már bőven egy százalék feletti javulás tapasztalható és mindez viszonylag gyorsan elérhető. Elég abba belegondolni, hogy a költségek esetén mindössze 3 munkatárssal és 200 óra munkával számoltam, amelynek eredményeként ennél jobb modellek is születhetnek és valószínűleg még idő is marad, de egy közepesebb nagyságú bérelt csapat esetében egy-két hónap alatt az egész call centeres (vagy akár nem csak a call centeres) értékesítés véletlen erdő alapúvá tehető, amely segítségével több százalékkal növelhető a vállalat éves működési nyeresége.

Az egyik leglényegesebb, a modell erejét legjobban szemléltető mutató a ROMI. A 365 528 488 várható bevétel növekedés és 230 268 652 költség csökkenésre, mindösszesen 47 406 124 kiadás jut, tehát a projekt egy éves ROMI-ja 1156%, ami kimagasló eredmény és egyértelműen a projekt megvalósítására ösztönöz, amennyiben nem találunk ennél magasabb egy éves ROMI-val rendelkezőt.

Érdekes összehasonlításba vetnünk az NPTB modell szakirodalma alatt említett ROI-val, amely összességében a direkt marketing kampányokra és nem kifejezetten a kereszt- és felülértékesítésre vonatkozott, ott azonban 530% feletti ROI-t mutatott. Az az érték is igen

magasnak számított, azonban jelen esetben láthatjuk, hogy ennél több, mint kétszer nagyobb értékről van szó, tehát a véletlen erdő, mint választott modell ezen a téren is sikeresnek mondható.

A kimagasló ROMI kérdéseket vethet fel bennünk, hogy a való életben is megjelenhetnek-e ezek a számok, erre saját tapasztalataim alapján is igen válasz adható. Alapfokú szakdolgozatom során még nem gépi tanulási módszerek, de adatelemzés alapján egy utazási iroda esetében alakítottam át a repülőjegy értékesítés esetén a kiegészítő szolgáltatások (ülőhelyfoglalás, feladott poggyász, elsőbbségi beszállás) árazását az adatok alapján. Az elemzés után az új árazás implementálásra is sor került, melynek hatására a kiegészítő szolgáltatásokból származó havi profit közel 100%-kal növekedett, melyhez költségként egyedül az én minimális egy havi (ennyi ideig tartott az a projekt) gyakornoki költségem társult.²⁸

Ezen gyakorlati tapasztalatom alapján a fenti kimagasló ROMI egyértelműen elérhető a véletlen erdő módszer segítségével.

VII. További előnyök:

Mindezek mellett több különböző nem, vagy csak részben számszerűsíthető eredményt is bemutatok:

Felvetődik egy részről a fogyasztói élmény javulása. A heurisztikus felhívásos módszerrel sokszor hívtak fel olyan ügyfeleket, akik igazából egyáltalán nem akartak, talán nem is gondoltak arra, hogy magasabb internet csomagra váltsanak és a mai személyre szabott világban nem vonzó egy olyan ajánlat, aki általánosságban kínál termékeket, olyan hatást keltve ezzel, hogy nem is foglalkoznak az adott fogyasztóval.

Ennek egy másik irányból való megközelítése az, hogy a fogyasztónak kevesebbet kell válogatnia bizonyos termékek közül, ismerkednie velük, összehasonlítani más vállalatokéval. Mindez történhet azért, mert a modell segítségével képesek lehetnénk előbb megtudni azt, hogy a fogyasztónak szüksége van-e rá, mint, hogy elkezdene érdeklődni iránta vagy adott esetben bejönne egy üzletünkbe. Ezzel időt és energiát spórol meg a fogyasztó, illetve azzal, hogy csökkentjük az üzletek látogatását és az ügyfélszolgálat vagy más platformok felkeresését érdeklődési céllal, a saját költségeinket is képesek vagyunk csökkenteni.

²⁸ <http://szd.lib.uni-corvinus.hu/12722/>

További előny származtatható az ügyfélszolgálat felszabaduló idejéből, amely a telesales hívások csökkenéséből származik. Ahogy korábban felvezettem, az ügyfélszolgálati partnerek számának redukálása helyett / mellett érdemes lehet átképezni / szerződést kötni olyan dolgozókkal, akik az ügyfélélményt képesek javítani, legfőképpen a panaszkezelés, hibabejelentés, illetve az érdeklődések esetén. Iparágtól független probléma, hogy sokszor várni kell arra, hogy kapcsoljanak valakit egy hibabejelentés esetén, amely az amúgy is frusztrált fogyasztót tovább frusztrálhatja, ezért számukra is nagy pozitív meglepetés lehet, ha egyből vagy nagyon rövid időn belül kapcsolhatóvá válik egy munkatárs. A panaszkezelés pedig szintén a szolgáltatások esetében egy kiemelkedő probléma, hiszen alapvetően a panaszok jelenléte nem csak átok, hanem áldás is lehet, hiszen tanulhat belőle a vállalat, viszont ehhez kellően jó panaszkezelési módszerre van szükség, amelyben szintén nagy előny lehet, ha több munkatárs képes fogadni és kezelni ezeket a panaszokat.

A munkatársak ezen oldalon való növekedése és a terhelés csökkenése a munkavállalók számára is pozitív hatással lehet, hiszen nem kell olyan nagy pörgéssel, gyorsasággal dolgozniuk, több idejük jut egy-egy ügyfél problémáját kellőképpen körbejárni.

A Telekom példa esetében a szemfüles olvasó észlelhette, hogy tréning költségekkel nem számoltam. Ennek oka az, hogy a modellt a gyakorlatban az ügyfélszolgálatosoknak nagyon könnyű alkalmazni, nem kell hozzá extra ismeret. A bevezetéshez mindösszesen az szükséges, hogy valaki lefuttassa (vagy automatán lefusson) a modellt minden reggel, majd pedig az adott napon migrálónak jelölt ügyfeleket jelölő lista felkerüljön a vállalat felhő alapú tárhelyébe vagy az ügyfélszolgálat által használt rendszerbe, vagy egy mindannyiuk által használt, együttesen szerkeszthető felületre. Itt egyértelműen látják, hogy kik azok, akiket érdemes felhívni, nincs szükség válogatásra, nem megy el rá extra idejük, hogy kigondolják, kik azok az ügyfelek, akiket adott nap érdemes felhívni.

Itt előnyként jelenik meg az IT részére az, hogy elég egyszer beállítani az adatbázist, ahonnan a modell az adatokat nyeri ki és onnantól az összes új hívással keletkező adat is bele fog kerülni, tehát egyre több adatból tanulhat a modell, melynek eredményeként egyre jobban és pontosabban fog működni minden valószínűség szerint. Itt érdemes lehet meghatározni bizonyos időintervallumokat- például utóbbi 2-3 év-, amelyet a modell vizsgál, ezzel jobban képes lehet kezelni az új trendeket, és a régiek nem fogják elfogulttá tenni a modellt.

A Telekom eset bemutatásakor nem tértem ki arra, hogy gyakorlatilag még ennél is több ügyfelet lehetne felkeresni és értékesíteni nekik, hiszen nem vettem számba a lemorzsolódást.

(Erre vonatkozóan a Telekomnak csak a mobilpiacra volt adata a dokumentumokban.) Az éves lemorzsolódási ráta a telekommunikációs szektorban 10 és 67 százalék között mozog²⁹, ami egy óriási szám és, ahogy az irodalomelemzés során előtérbe került, az ügyfélmegtartás növelése nagy mértékben növelheti a profitot és az értékesítés véletlen erdőre való helyezése is elősegítheti ezt. Egyrészt magával azzal, hogy több ügyfelet lehet felhívni, lehetséges, hogy képesek olyan ügyfeleket felhívni és nekik tetsző ajánlatokat tenni, akik amúgy lemorzsolódtak volna, ezzel pedig csökken a migráció, nő az értékesítés, a vevőélettartam érték és összességében a bevételek és profit is.

Ez még tovább fejleszthető azzal, hogy létrehozunk egy véletlen erdő modellt, amely segítségével a migráció helyett a lemorzsolódó ügyfeleket lehetne előre jelezni. Amennyiben ezeket az ügyfeleket sikeresen beazonosítjuk, akkor lényegesen könnyebb dolgunk van elérni a megtartásukat, az ő egyedi eseteikben, mint ha minden ügyfelet kívánnánk egységes módon megtartani.

A véletlen erdő segítségével a valóságban az értékesítési számok még ennél is nagyobb mértékben nőhetnek, mivel az adatbázisban csak az látszódtott, hogy milyen értékesítés történt, az nem került rögzítésre, hogy mely termékek kerültek felajánlásra. Lehetséges, hogy a modell esetében a tréning alatt fals pozitívnak jelölt ügyfelek egy része tényleg magasabb internet előfizetésre váltott volna, de mi TV előfizetést ajánlottunk neki.

VIII. Legnagyobb kihívások és a modell korlátai

VIII.1. Legnagyobb kihívások

A modell felépítése során több olyan kihívás is felmerült, amelyekkel meg kellett küzdenem. Az egyik kihívás maga a véletlen erdő módszer kiismerése volt. Döntési fákkal dolgoztam már tanulmányaim során, azonban a véletlen erdő módszert és részletesebben a gépi tanuló algoritmusok világát is ezen tanulmány miatt ismertem meg, emiatt sokat kellett különböző szakmai fórumokon és blogokon keresnem mások példáit, útmutatóit megtekintve. Azonban azt is megismertem eközben, hogy a nap mint nap sok-sok éve ebben dolgozó adattudósok is ezt teszik, hiszen a technológia napról napra fejlődik, nyílnak meg új lehetőségek, kerülnek fel jobbnál jobb útmutatók, amelyek alapján az adattudós fejlesztheti modelljeit és átlendülhet a kihívásain.

²⁹ <http://www.dbmarketing.com/telecom/churnreduction.html>

Egy másik nagy kihívás volt a rengeteg jellemző értelmezése. Egyrészt fontos volt tudni, hogy mely jellemzők pontosan mit jelentenek, vannak jellemzők, ahol a név nem feltétlen utal arra, hogy gyakorlatban mire vonatkoztat a modell, máshol pedig a mérési egység nem egyértelmű. Ezeket fontos a feladat legelején tisztázni és pontosan utána járni, és mindenképpen elengedhetetlen dokumentálni ezeket, majd pedig ellenőriztetni is őket az illetékesekkel, illetve adott esetben a rendszerekkel, ahonnan az adatok származnak, hogy minden helyesen szerepel-e. Az én modellem alkotásának a végén derült ki, hogy félreértések miatt az előző hívás eredményét prediktáltam és nem az aktuális hívását. Ezután nem kellett mindent az elejéről kezdeni, de idővesztéssel járt, amely egy éles szituációban, adott esetben megrendelésen alapuló vállalati projekt esetén pénzügyi veszteségekkel is járhat.

Ezek mellett a menedzsment, az informatika és az értékesítés oldaláról való támogatás megszerzése, kiépítése a legnagyobb kihívás véleményem szerint. Ez annyiban kakukk tojás, hogy nem teljes mértékben a modell építése során derült ki, hiszen korábbi tanulmányaim és munkám során is előkerült, hogy az informatikához köthető projektek esetében a menedzsment támogatásának az elnyerése az egyik legfontosabb feladat. Amennyiben a menedzsment támogatja feladatot, az sokkal egyszerűbben és gyorsabban lehet sikeresen és gyorsabban is lesz elvégezhető. A menedzsment bizalmának eléréséhez fontos, hogy megfogalmazzuk számukra is könnyen érthetően, tömören, de bizonyíthatóan, hogy miért éri meg a cégnek az, amit csinálni kívánunk. Amennyiben ezt megteesszük, valószínűleg megkapjuk a szükséges bizalmat.

Ezen modell esetén a menedzsmenten kívül azonban két másik oldalt is számításba kell venni. Egy részről az informatikáét, ez egy nagyobb szervezet esetén különösen lényeges lehet, ahol különböző osztályok különböző adatbázisait kell megszerezni - itt legfőképpen adatbiztonsági kérdések vetődhetnek fel, ahol senki nem kívánja kiadni saját adatait. Fontos azt a fajta bizalmat kiépíteni, hogy az adatok biztonságban lesznek kezelve, ezt legfőképpen különböző szabványoknak megfelelő adatkezelési módokkal és szigorú szerződésekkel lehet elérni, de ebben segít például a KNIME Server legdrágább verziója is, ahol a modell esetében is van beépített titkosítás.

Végül de nem utolsó sorban pedig az értékesítési oldalt is. Itt azon vállalatok, akiknek a telesales csapatai ki vannak szervezve előnyben lehetnek azon a téren, hogy nem vagy kevésbé kell tartaniuk attól, hogy a dolgozók körében esetlegesen „géprombolás” következik be. Géprombolás történhet amiatt, mert úgy gondolják az alkalmazottak, hogy a mesterséges

intelligencia elveszi a munkájukat, emiatt nem hajlandóak azt használni, fellépnek a bevezetése ellen vagy éppen elhagyják a vállalatot. Amennyiben partner cégekről van szó, ez kevésbé történhet meg, hiszen egy-egy partner adott esetben nem csak egy vállalattal van szerződésben, így könnyebben átszervezésre kerülhet a munkaerő, mint a vállalaton belüli ügyfélszolgálat esetén. Azonban meg kell jegyezni azt, hogy az értékesítési oldal tudtára legyen adva, hogy a modell nem ki akarja váltani az ügyfélszolgálatot, hanem effektívebbé tenni azt, hiszen, ahogy a szakirodalmi áttekintés alatt kitértem rá: egy gép-ember kombináció általában jobban képes működni, mint gép-gép vagy ember-ember társai.

Ezt elősegíti a megfelelő, újításokra nyitott vállalati kultúra kiépítése, amelyben a munkavállaló ugyanúgy érték marad, csak a munkaköre (például innentől panaszkezeléssel foglalkozik) vagy a munkája elvégzésének a módja változik olyan mértékben, hogy a szabadság érzete is megmaradjon. Fontos garantálni, hogy ténylegesen azokat a fogyasztókat keressék, meg akiket a modell ajánl (ez azzal könnyen elérhető, hogy csak a modell által ajánlott fogyasztók listáját látjuk), viszont azt, hogy őket például milyen sorrendben hívják fel, már nem szükséges és nem is szabad megkötni, hiszen ezzel túlságosan nagy mértékben kerülnének korlátozásra.

VIII.2. A modell korlátai és fejlesztési lehetőségek

Tökéletes gépi tanulási módszer vagy modell nem létezik, így ennek a modellnek is megvannak a hiányosságai és korlátai. „*Minden modell olyan jó, mint az adat, amivel dolgozik*” szokták mondani. Az én modellem esetében nagy előny volt, hogy rendelkezésre állt rengeteg adat, egy helyre gyűjtve, fogyasztható formában. Amennyiben nekem kellett volna különböző osztályoktól begyűjteni az adatokat, majd egy sémára rendezni őket (például megállapítani különböző Facebook üzenetekről, vagy javítási visszajelzésről, hogy melyik volt negatív vagy pozitív hangvételű), az lényegesen több időt vett volna igénybe. Emiatt is számoltam az implementációnál több idővel és több tapasztalt dolgozóval, ügyelve arra, hogy bizonyára nem lesz olyan egyszerű dolguk, mint nekem, illetve az informatikai osztály támogatása is rendkívül nagy segítséget biztosít ezen a téren.

Ezzel valamelyest ellent mondva a rendelkezésre álló adatok korlátoztak is valamelyest. Jelenleg a modell azt nem tudja meghatározni, hogy adott ügyfélnek pontosan melyik csomagot vagy csomagokat célszerű ajánlani, mivel a rendelkezésre álló adatok között nem volt olyan jellemzőnk, amely megmondta volna, hogy a vásárlás esetén melyik termékre váltott pontosan az ügyfél. Ezen adatok minden valószínűség szerint gyűjtésre kerülnek valahol és célszerű is

őket gyűjteni, hiszen ennek a segítségével az értékesítési folyamat még jobban gyorsíthatóvá és személyre szabhatóvá válik, továbbá jobb és pontosabb eredményeket is kaphatunk (illetve a teszt adatsoron kapott eredményeink is jobban validálhatóvá válnak, ahogy erre korábban utaltam). Szintén hiányoztak azok a jellemzők, amelyek meghatározták, hogy pontosan mely termékek vagy terméktípusok voltak felajánlva az ügyfeleknél bizonyos hívásoknál. Emiatt csak körülbelül voltam képes meghatározni, hogy a hívások hány százalékában ajánlottak konkrétan magasabb internet terméket az előfizető szám és az összes értékesítési opció alapján. Amennyiben ezt rögzítjük, akkor pontosabb becsléseket lesz képes adni a modell és az eredményességet is jobban nyomon tudjuk követni. Példaként megemlíthetők még az általam már korábban említett versenytársi árak. Jelenleg tudjuk, hogy melyek azok az ügyfeleink, akik olyan területeken laknak, ahol más szolgáltató is jelen van és azt is rögzítjük, hogy melyek ezek a (nagyobb) szolgáltatók, azonban azt nem, hogy milyen árakkal dolgoznak. Amennyiben ezeket rögzítjük szintén javulhatnak a modell eredményei. ezek csak példák voltak jelen esetünkben, azonban ezek mind elősegíthetik az értékesítési számok növekedését és a hívások mennyiségének optimalizálását.

A modell alkotása során voltak olyan adattudományhoz köthető lépések és területek, amelyeket nem, vagy csak részlegesen érintettem. Ennek két oka volt. Egyrészt a géperő: voltak olyan módszerek, melyeket ki akartam próbálni, de egyszerűen olyan lassú lefutása volt (itt várhatóan minimum órákra kell gondolni), hogy végül elvettem őket. Ez komolyabb szervergépekkel vagy felhő alapú kalkulációkkal orvosolható és ajánlott is a nemzetközi legjobb módszerek alapján az általam említett módszereken kívül másokat is kipróbálni. Ezen okból került elvetésre a modell esetében az adatok kiegyensúlyozására szolgáló SMOTE módszer, az ügyfélszintű kiértékelést, vagy a modell érthetőségét elősegítő módszerek közül a LIME és a SHAP.

A másik ok a rendelkezésre álló időkeret volt. Ennek hiányában nem került a feature engineeringre (azaz további jellemzők létrehozására a meglévő jellemzők alapján pl. korból korcsoportok, egyesek szerint a feature engineering maga az alkalmazott gépi tanulás³⁰), a H2O.ai platformon (egy különálló KNIME-hoz hasonló, ám KNIME integrációval rendelkező platform, amelyen rendkívül sok modell építési, tervezési mód automatizált) való modell építésre és az olyan további módszerek hasznosítására, amelyekhez különböző programkódokat

³⁰ <https://towardsdatascience.com/importance-of-feature-engineering-methods-73e4c41ae5a3>

kellett volna az esethez hangolnom vagy megírnom, mint például a MissForest algoritmus, mely egy szintén véletlen erdő alapú módszer a hiányzó értékek kezelésére.

Potenciális veszélyként jelenik meg a kannibalizáció. Fennáll a lehetőség arra, hogy olyan fogyasztókat keres fel az értékesítés, akik a megkeresésünk nélkül is magasabb csomagra váltanának. ez jelenleg a heurisztikus értékesítés során is fenn álló probléma és felesleges költségekkel terheli a vállalatot, azonban azzal, hogy több potenciális vásárlót találunk meg, nő azoknak a száma is, akiket „feleslegesen” keresünk fel. Ennek orvoslására fontos meghatározni a jelenlegi kannibalizációs arányt, illetve, hogy az hogyan változik a modell kiépítése után. Amennyiben ez az arány magas, érdemes lehet a modellen fejleszteni vagy egy külön modellt létrehozni arra, hogy elkülöníthessük azokat, akik valószínűleg a megkeresésünk nélkül is migrálnának. Ehhez érdemes összehasonlítani a telesales adatokat azokkal az értékesítésekkel, akik maguk keresték meg az ügyfélszolgálatot azzal, hogy csomagot váltanának.

Korábban említésre került, hogy az EDM kampányok értékesítéseire vonatkozóan is álltak rendelkezésemre adatok, ám pontosan az, hogy ki volt, aki esetleg a linkre kattintás után váltott magasabb csomagra egy adott időintervallumon belül, nincsen meghatározva. Ez azért lényeges, mert az EDM kampányok általánosságban olcsóbbak, mint a telefonos megkeresések, ezért érdemes lehet egy modell segítségével megállapítani nem csak azt, hogy ki az, aki magától váltana, hanem azokat is, akiknek adott esetben elég e-mailt küldeni, nem szükséges a telefonos felkeresés. Adott esetben az is lehetséges, hogy e-mailre kedvezőbben reagálnak bizonyos ügyfelek, mint a telefonhívásra.

A modell létrehozása után a munka nem ér véget, a karbantartással is foglalkozni kell. Erre jelen modellalkotás közben nem került sor, mivel a modell éles bevezetésre nem került azonban a fejlesztésnek szerves része. Ahogy korábban említettem, amennyiben változik a technológia vagy akár a gazdasági környezet, különböző átalakulások történnek a társadalomban, úgy a jellemzőink is változnak velük együtt (például a Facebook megjelenése, és ezzel együtt az oda tolódó panaszbejelentések és azok kezelésének és dokumentálásának jelentősége megnőtt). A véletlen erdő nagy előnye, hogy amennyiben valamilyen apró változás történik a jellemzőkön belül, akkor a modell képes azt észlelni és a hatásait is felméri, azaz képes akár egyetlen új lefuttatással átszervezni a saját belső működését, amelyet a felhasználók nem is észlelnek, mivel a lefuttatási idő érdemben nem változik.

Azonban a modell olyan változásokat nem képes kezelni, amelyek nincsenek a jellemzőink között. Így amennyiben a modell hirtelen romlásnak indul vagy még jól működik,

de úgy érezzük, hogy valamilyen változás történt, amelyhez szükséges új jellemzőt felvennünk, akkor azt érdemes megtenni minél előbb, hiszen amennyiben a jellemző jelentősége akkor még alacsony, azt a modell fel fogja ismerni és ezt megfelelően fogja tudni kezelni, ugyanúgy azt is, ha később megnőne a jelentősége. Szintén a karbantartáshoz köthető a korábban említett időintervallum használata: azaz például kifejezetten az utóbbi egy-két évben képződött hívások adataival dolgozni, itt a pontos időintervallum kiválasztásához érdemes A / B tesztet folytatni és ez alapján dönteni, a teszteléshez ötletek meríthetők nemzetközi legjobb módszerekből, illetve vállalaton belül gyakran használt intervallumok használata is tesztelhető.

VIII.3. A Pontosság és a Felidézés érzékenységeinek összehasonlítása

Mindezeket mindenképpen érdemes számba venni és törekedni arra, hogy a modellünk tovább javuljon, mivel kicsinek tűnő változtatások is nagyon nagy hatással lehetnek a projekt sikerességére és ezáltal a vállalatunk eredményére. A Magyar Telekomos esetet felidézve az alábbi két táblázatban láthatjuk, hogy bizonyos esetekben mi történik a Pontosság és a Felidézés változása esetén.

Felidézés:	Profit változás (HUF):
89,2%	637 000 321
70,0%	597 473 080
55,0%	547 391 016
55 > 89,2%	89 609 306

21. Táblázat A Felidézés változásának hatása a profitra.

Forrás: Saját szerkesztés.

A Felidézés esetében azt tapasztalhatjuk, hogy a modell eredménye és a legpesszimistább eset között ~ 90 millió forint eltérés tapasztalható vagy másképp fogalmazva 24,2 százalékpontnyi változás a Felidézésben ~16,4%-kal nagyobb profít növekedést jelent a jelenlegi állapothoz képest. Ez a szám önmagában jelentős, azonban érdemes összehasonlítani azzal, hogy mi történik a Pontosságban végbemenő változások esetén. Itt érdemes megjegyezni, hogy nincsenek számításba véve a felszabaduló idővel járó magasabb minőségű panaszkezelésből vagy hibabejelentések kivizsgálásából járó nehezen számszerűsíthető bevételek, amelyek a magasabb Felidézés esetén további profitot jelenthetnek, ezáltal a különbség nagyobb is lehet a gyakorlatban.

Pontosság:	Profit változás (HUF):
72,0%	-22 180 532
79,9%	547 391 016
80,0%	554 583 372
80,9%	619 471 944
81,9%	691 569 727
79,9 > 81,9%	144 178 712

22. Táblázat A Pontosság változásának hatása a profitra.

Forrás: Saját szerkesztés.

A Pontosság esetén azt kell megállapítanunk, hogy egy százalékpontnyi változásnak lényegesen nagyobb hatása van a profitra, mint a Felidézés esetén. Ahogy a táblázatból láthatjuk (minden érték esetében 55%-os Felidézéssel számolva), ha a 79,9%-os precízió már 0,1 százalékponttal javul, akkor is több millió forint extra hasznot könyvelhet el a vállalat, azonban ami jól szemlélteti a különbségeket a Felidézéshez képest az a 2 százalékpontnyi növekedés esetén történő ~144 millió forintos profitnövekedés, ami ~ 60%-kal nagyobb, mint a Felidézés esetén bekövetkező 24,2 százalékpontnyi növekedés.

Mindezek tudatában elmondható, hogy ezáltal sikerült elérni mindkét kitűzött célt: az értékesítések számának a növelését és a hívásszám csökkentését is. A Felidézés növelése mindenképpen fontos feladat, hiszen ezzel voltunk képesek elérni azt, hogy több potenciális ügyfelet kereshessünk meg egy évben, azonban a modellt jelen esetben érdemes kiemelten a precízió javítására összpontosítani hiszen már egy százalék pontnyi növekedés is rendkívül magas pénzügyi haszonnal kecsegtethet. Azonban óvatosnak is kell lenni, hiszen a 72%-os precízió hirtelen jónak tűnhet, azonban mint a táblázatban láthatjuk, ez már kevesebb profitot hozna az 55%-os Felidézés mellett, mint a kiinduló heurisztikus állapot.

IX. Összefoglalás:

Tanulmányom kutatási kérdése a következő volt: „*a véletlen erdő módszer segítségével milyen módon és mekkora mértékben segíthető elő a hazai telekommunikációs szektorban a felül- és keresztértékesítés?*”.

Dolgozatom elején alapos szakirodalmi áttekintés segítségével bemutattam a témához kötődő legfontosabb fogalmakat, az ezekhez kapcsolódó trendeket és a telekommunikációs, illetve pénzügyi szektorból származó nemzetközi kutatásokat. Ezután három előrejelzéshez alkalmazható módszert mutattam be. A marketingben jól ismert és használt logisztikus

regresszió mellett a kevésbé ismert, de gyakorlatban bizonyított NPTB módszer (Knott, 2002) is bemutatásra került, végül az általam is használt véletlen erdő modellt szemléltettem.

Ezt követően bemutattam a valós vállalati adatokon végzett kutatásomat, melynek során egy véletlen erdő modellt építettem. Ügyeltem a legmagasabb részletességre, hiszen célom volt az, hogy az olvasók saját szervezetükben is implementálni tudják a modellt.

A modell a teszt adatokon 79,9%-ban képes volt megtalálni a magasabb internet csomagra váltó fogyasztókat, illetve képes volt elérni, hogy a hívások 89,2%-a sikeressé váljon, ellentétben az eddigi körülbelül 26,16%-os sikerességgel. Ezek közül kifejezetten a hívások sikerességi arányának növelését tartottam nagy eredménynek, amely miatt pesszimistább értékekkel is végeztem a későbbiekben teszteléseket a példa eseten. Fontos megállapítássá vált, hogy a felszabaduló idő segítségével olyan fogyasztókat is elérhetünk, akikre eddig nem volt idő.

Kutatásomból kiderült az is, hogy az eredetileg egyenlő fontossággal kezelt Pontosság és Felidézés értékek közül előbbi adott százalékos változása van nagyobb hatással a vállalat pénzügyi eredményeire. Itt állapítottam meg, hogy az elsőre akár megfelelőnek tűnő 72%-os Pontosság még negatív megtérüléssel jár, másik oldalról viszont egy nagyvállalat esetében már 0,1 százalékpontos javulás is további több millió forintos nyereségtöbbletet jelenthet.

Kitértem a legfontosabb és a legkevésbé fontos vágó kritériumok szemléltetésére, mely jó fényt vetített arra, hogy a modell milyen széles skálán gondolkodik és értékel. Többek között kiderült, hogy azt, hogy egyes területeken mennyire vannak jelen versenytársaink, illetve az, hogy a különböző elektronikus direkt marketing tevékenységeink közül mennyit nyitottak meg a fogyasztók, nem volt jelentős hatással a felülrértékesítésre az esetleges várakozásainkkal szemben. Azonban a legfontosabb vágó kritériumok esetén olyan nem várt jellemzőket találtunk, mint például, hogy valaki idős férfi-e vagy sem, illetve, hogy milyen TV csomaggal rendelkezik.

Ezután az eredmények kontextusba helyezésének érdekében felépítettem egy a Magyar Telekom Nyrt. pénzügyi adataira és a modell eredményeire épülő esetet, melynek a végén még a legpesszimistább becslés alapján is megállapítottam, hogy több, mint 1100%-os éves ROMI-val rendelkezik a projekt, amely az elméletben említett NPTB módszer ROI-jának több, mint kétszerese.

A modell további, számokban nem vagy kevésbé kifejezhető előnyeinek bemutatásakor kitértem a fogyasztói élmény javítására, illetve a munkavállalók eredményességének növelésére is. Felvázoltam a legnagyobb kihívásokat, mint a jellemzők értelmezése, vagy a vállalat különböző osztályai támogatásának megszerzése. Kitértem a modell korlátaira is: miért volt a rendelkezésre álló adatbázis előnyös és hátrányos is egyben, hogy milyen lépések voltak, amelyekre én a modell fejlesztése során nem tértem, de javíthatják a modell eredményességet. Felhívtam a figyelmet továbbá a karbantartás fontosságára és kannibalizáció kérdésére is, mint az implementáció során mindenképpen további vizsgálatokat igénylő tényezők.

Összességében tehát kutatásom segítségével képet kaphattunk arról, hogy a véletlen erdő modell a kereszt- és felülrértékesítés területén milyen széles területen képes előnyöket nyújtani és teszi mindezt úgy, hogy a hívások számát képes még a legpresszimistább esetben is közel 50%-kal csökkenteni, miközben az értékesítés is növekszik és ezáltal a véletlen erdőhöz kapcsolódó projektek ROI-ja 1000% fölé tehető.

Dolgozatommal betekinthezt az olvasó az adatvezérelt felülrértékesítés világába, előnyeibe egy gyakorlati és egy fiktív példa mentén. Az adatvezérelt marketing segítségével képes volt a modell olyan komplex összefüggések feltérképezésére, melyek a hagyományos marketing analitika által kínált megoldás számára rejtve maradhattak. Ennél többet is nyújt azonban a gépi tanulás, hiszen képes kezelni az adatokban történő legkisebb változásokat is, ezáltal pedig automatikusan követni a fogyasztóink, célcsoportjaink különböző jellemzőiben történő változásokat, amelyeket egy hipotézisekre felépített marketingtervezet nem képes ilyen rugalmassággal kezelni. Kutatásom alapján a marketingben eddig megszokott és használatos kvantitatív technikák kiváltására, modernizálásra alkalmas és a közel jövőben széles körben elterjedhet ezen a területen is a gép tanuló algoritmusok és ezáltal a véletlen erdők használata.

Azonban mivel a mesterséges intelligencia sok esetben fekete dobozként működik (habár vannak már kifejezetten a mesterséges intelligencia megértését segítő algoritmusok is), ezért amennyiben mélyrehatóan kívánjuk elemezni fogyasztóink hozzáállását bizonyos termékekhez, abban az esetben még mindig releváns és ajánlott különböző kvalitatív módszerek, például mélyinterjúk vagy fókuszcsoportok használata, amelyeket ötvözve a gépi tanulás erejével érhetjük el a legjobb eredményeket, bizonyítva ezzel azt, hogy az ember-gép kapcsolat ezen a területen is képes jobban teljesíteni mint a hagyományos ember-ember vagy az utópisztikus gép-gép páros.

A téma során rengeteg érdekes kérdéskörrel találkoztam, ezáltal és az előző bekezdésben is írtak alapján a továbbiakra az alábbi kutatási lehetőségeket javaslom:

- Marketingkutatás a XXI. században hibrid (mesterséges intelligencia és kvalitatív technikák) módszerrel
- A marketing költségvetés elosztásának optimalizálása mesterséges intelligencia segítségével

A dolgozatomat a kezdetéhez hasonlóan egy idézethez kapcsolódó gondolattal zárom. A marketing területén rendszeresen előkerül John Wanamaker hírhedt mondata³¹, miszerint: „*A reklámozásra elköltött pénzem fele a szemétként landol, a probléma az, hogy nem tudom, hogy melyik fele.*”. Az idő előrehaladtával egyre jobb módszerek jöttek létre annak elérése, hogy ez ne legyen igaz, azonban a kereszt- és felülrértékesítés területén, ahol sok helyen még mindig pusztán heurisztikus megoldásokkal dolgozunk, a gépi tanuló modellek képesek elhozni a változást.

³¹<https://www.b2bmarketing.net/en-gb/resources/blog/half-money-i-spend-advertising-wasted-trouble-i-dont-know-which-half>

X. Irodalomjegyzék

Agárdi, I. (2017). *Kereskedelmi marketing és menedzsment*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj233kmem_83_p2#dj233kmem_83_p2 [Hozzáférés dátuma: 2021.01.13.].

Agárdi, I. and Gyulavári, T. (2017). CRM-stratégiák a bank- és biztosítási szektorban. *Vezetéstudomány / Budapest Management Review*, 48(2).

Agrawal, A. (2017). *Logistic Regression. Simplified*. [online] Medium. Elérés helye: <https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389> [Hozzáférés dátuma: 2021.01.16.].

Ahn, H., Song, C., Ahn, J.J., Lee, H.Y., Kim, T.Y. and Oh, K.J. (2010). Using Hybrid Data Mining Techniques for Facilitating Cross-Selling of a Mobile Telecom Market to Develop Customer Classification Model. *2010 43rd Hawaii International Conference on System Sciences*.

Ahn, J.-H., Han, S.-P. and Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 30(10-11), pp.552–568. doi:10.1016/j.telpol.2006.09.006.

Akçura, M.T. and Srinivasan, K. (2005). Research Note: Customer Intimacy and Cross-Selling Strategy. *Management Science*, 51(6), pp.1007–1012.

Bauer, A. and Kolos, K. (2017). *Márkamenedzsment*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj201mm_43_p1#dj201mm_43_p1 [Hozzáférés dátuma: 2021.01.13.].

Bernschütz, M., Deés, S. and Kenéz, A. (2016). *Marketing esettanulmányok*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj75me_68_p3#dj75me_68_p3 [Hozzáférés dátuma: 2021.01.14.].

Bhatti, H.S., Abareshi, A. and Pittayachawan, S. (2016). An Empirical Examination of Customer Retention in Mobile Telecommunication Services in Australia. *Proceedings of the*

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), pp.197–227.

Bose, I. and Mahapatra, R.K. (2001). Business data mining — a machine learning perspective. *Information & Management*, [online] 39(3), pp.211–225. Elérés helye: <https://www.sciencedirect.com/science/article/pii/S037872060100091X> [Hozzáférés dátuma: 2021.01.13.].

Chen, C.-F. and Cheng, L.-T. (2012). A study on mobile phone service loyalty in Taiwan. *Total Quality Management & Business Excellence*, 23(7-8), pp.807–819.

Chowdary, D.H. (2020). *Decision Trees Explained With a Practical Example*. [online] Medium. Elérés helye: <https://medium.com/towards-artificial-intelligence/decision-trees-explained-with-a-practical-example-fe47872d3b53> [Hozzáférés dátuma: 2021.01.16.].

Cohen, D., University, L. and Al, E. (2006). *Customer satisfaction : a study of bank customer retention in New Zealand*. Christchurch, N.Z.: Commerce Division, Lincoln University.

Danyi, P., Rekettye, G. and Veres, I. (2020). *Modern árazás*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/m770modar_22_p4#m770modar_22_p4 [Hozzáférés dátuma: 2021.01.13.].

Darzi, M.A. and Bhat, S.A. (2018). Personnel capability and customer satisfaction as predictors of customer retention in the banking sector. *International Journal of Bank Marketing*, 36(4), pp.663–679.

Denizci Guillet, B. (2020). Online upselling: Moving beyond offline upselling in the hotel industry. *International Journal of Hospitality Management*, 84, p.102322.

Eichinger, F., Nauck, D.D. and Klawonn, F. (2006). Sequence Mining for Customer Behaviour Predictions in Telecommunications. *ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges*.

Evanschitzky, H., Ramaseshan, B., Woisetschlager, D.M., Richelsen, V., Blut, M. and Backhaus, C. (2011). Consequences of customer loyalty to the loyalty program and to the company. *Journal of the Academy of Marketing Science*, 40(5), pp.625–638.

Gould, G. (1995). Why it is customer loyalty that counts (and how to measure it). *Managing Service Quality: An International Journal*, 5(1), pp.15–19.

Gyulavári, T., Mitev Ariel, Z., Neulinger, Á., Neumann-Bódi, E., Simon, J. and Szűcs, K. (2017). *A marketingkutatás alapjai*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj240ama_55_p1#dj240ama_55_p1 [Hozzáférés dátuma: 2021.01.15.].

Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10), pp.2902–2917.

Heidrich, B. (2017). *Szolgáltatás menedzsment*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj252szm_97_p4#dj252szm_97_p4 [Hozzáférés dátuma: 2021.01.13.].

Horváth, D. and Bauer, A. (2016). *Marketingkommunikáció*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj58m_26_p6#dj58m_26_p6 [Hozzáférés dátuma: 2021.01.12.].

Johnson, J.S. and Friend, S.B. (2014). Contingent cross-selling and up-selling relationships with performance and job satisfaction: an MOA-theoretic examination. *Journal of Personal Selling & Sales Management*, 35(1), pp.51–71.

Kacsukné Bruckner, L. and Kiss, T. (2019). *Bevezetés az üzleti informatikába*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/m596bau_135_p1#m596bau_135_p1 [Hozzáférés dátuma: 2021.01.15.].

Kamakura, W.A., Kossar, B.S. and Wedel, M. (2004). Identifying Innovators for the Cross-Selling of New Products. *Management Science*, 50(8), pp.1120–1133.

Kamakura, W.A., Wedel, M., de Rosa, F. and Mazzon, J.A. (2003). Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*, 20(1), pp.45–65.

Kaur, H. and Soch, H. (2012). Validating Antecedents of Customer Loyalty for Indian

Cell Phone Users. *Vikalpa: The Journal for Decision Makers*, 37(4), pp.47–62.

Keiningham, T.L., Aksoy, L., Williams, L. and Buoye, A. (2011). Why Loyalty Matters in Retailing. *Service Management*, pp.67–82.

Keller, K.L. and Kotler, P. (2016). *Marketingmenedzsment*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj183m_132_p30#dj183m_132_p30 [Hozzáférés dátuma: 2021.01.13.].

Kemény, I. (2015). *A versenytársak csak egy kattintásra vannak. Az újravásárlási és továbbajánlási szándék alakulása különböző e-kereskedelmi kategóriákban*. Budapest: Corvinus University of Budapest.

Kenesei, Z. and Cserdi, Z.H. (2018). *Szolgáltatásmarketing a turizmusban*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj319szmat_18_p1#dj319szmat_18_p1 [Hozzáférés dátuma: 2021.01.13.].

Kenesei, Z. and Kolos, K. (2014). *Szolgáltatásmarketing és -menedzsment*. Budapest: Alinea.

Knott, A., Hayes, A. and Neslin, S.A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing*, 16(3), pp.59–75.

Li, Q. (2009). An Algorithm of Quantitative Association Rule on Fuzzy Clustering with Application to Cross-Selling in Telecom Industry. *2009 International Joint Conference on Computational Sciences and Optimization*, 1.

Li, S., Sun, B. and Montgomery, A.L. (2011). Cross-Selling the Right Product to the Right Customer at the Right Time. *Journal of Marketing Research*, 48(4), pp.683–700.

Liu, L. (2008). Study of the Relationship Between Customer Satisfaction and Loyalty in Telecom Enterprise. *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*.

Liu, T.-C. and Wu, L.-W. (2007). Customer retention and cross-buying in the banking industry: An integration of service attributes, satisfaction and trust. *Journal of Financial*

Services Marketing, 12(2), pp.132–145.

Malhotra, N.K. and Simon, J. (2017). *Marketingkutatás*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj198mk_1_p1#dj198mk_1_p1 [Hozzáférés dátuma: 2021.01.16.].

Mohd Kassim, N. and Souiden, N. (2007). Customer retention measurement in the UAE banking sector. *Journal of Financial Services Marketing*, 11(3), pp.217–228.

Mutanen, T., Ahola, J. and Nousiainen, S. (2006). Customer churn prediction - a case study in retail banking. *ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges*.

Olaru, C. and Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, 138(2), pp.221–254.

Oshiro, T.M., Perez, P.S. and Baranauskas, J.A. (2012). How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, 7376, pp.154–168.

Piskóti, I. (2016). *Businessmarketing-menedzsment*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj106bm_39_p20#dj106bm_39_p20 [Hozzáférés dátuma: 2021.01.13.].

Poór, J. (2017). *Menedzsment-tanácsadási kézikönyv*. [online] Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj212mtkk_176_p3#dj212mtkk_176_p3 [Hozzáférés dátuma: 2021.01.14.].

Prinzie, A. and Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems*, 44(1), pp.28–45.

Qi, J., Wu, F., Li, L. and Shu, H. (2007). Artificial intelligence applications in the telecommunications industry. *Expert Systems*, 24(4), pp.271–291.

Reichheld, F.F. and Sasser, W.E.Jr. (1990). Zero defections: quality comes to services. *Harvard Business Review*, September–October.

Rekettye, G., Törőcsik, M. and Hetesi, E. (2016). *Bevezetés a marketingbe*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj81bam_93_p13#dj81bam_93_p13 [Hozzáférés dátuma: 2021.01.12.].

Sanjeevi, M. (2017). *Chapter 4: Decision Trees Algorithms*. [online] Medium. Elérés helye: <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1> [Hozzáférés dátuma: 2021.01.16.].

Schmitz, C., Lee, Y.-C. and Lilien, G.L. (2014). Cross-Selling Performance in Complex Selling Contexts: An Examination of Supervisory- and Compensation-Based Controls. *Journal of Marketing*, 78(3), pp.1–19.

Swinhoe, D. (2018). *The 18 biggest data breaches of the 21st century*. [online] CSO Online. Elérés helye: <https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html> [Hozzáférés dátuma: 2021.01.14.].

Szőke, P. (2015). Azok a mai fiatalok – Generációs különbségek a biztosítási szokásokban. *Biztosítás és Kockázat*, 2(3), pp.30–47.

Tayan, B. (2019). *The Wells Fargo Cross-Selling Scandal*. [online] papers.ssrn.com. Elérés helye: <https://ssrn.com/abstract=2879102> [Hozzáférés dátuma: 2021.01.16.].

Veres, Z. (2017). *A szolgáltatásmarketing alapkönyve*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj253asza_55_p15#dj253asza_55_p15 [Hozzáférés dátuma: 2021.01.13.].

Veres, Z., Hoffmann, M. and Kozák, Á. (2017). *Bevezetés a piackutatásba*. [online] Akadémiai Kiadó. Akadémiai Kiadó. Elérés helye: https://mersz.hu/hivatkozas/dj219bap_16_p1#dj219bap_16_p1 [Hozzáférés dátuma: 2021.01.15.].

Wharton (2018). *AI and Humans Working Together Bring Big Productivity Gains*. [online] Knowledge@Wharton. Elérés helye: <https://knowledge.wharton.upenn.edu/article/reimagining-work-age-ai/> [Hozzáférés dátuma: 2021.01.14.].

Yiu, T. (2019a). *Understanding Logistic Regression*. [online] towards data science. Elérés helye: <https://towardsdatascience.com/understanding-logistic-regression-using-a-simple-example-163de52ea900> [Hozzáférés dátuma: 2021.01.16.].

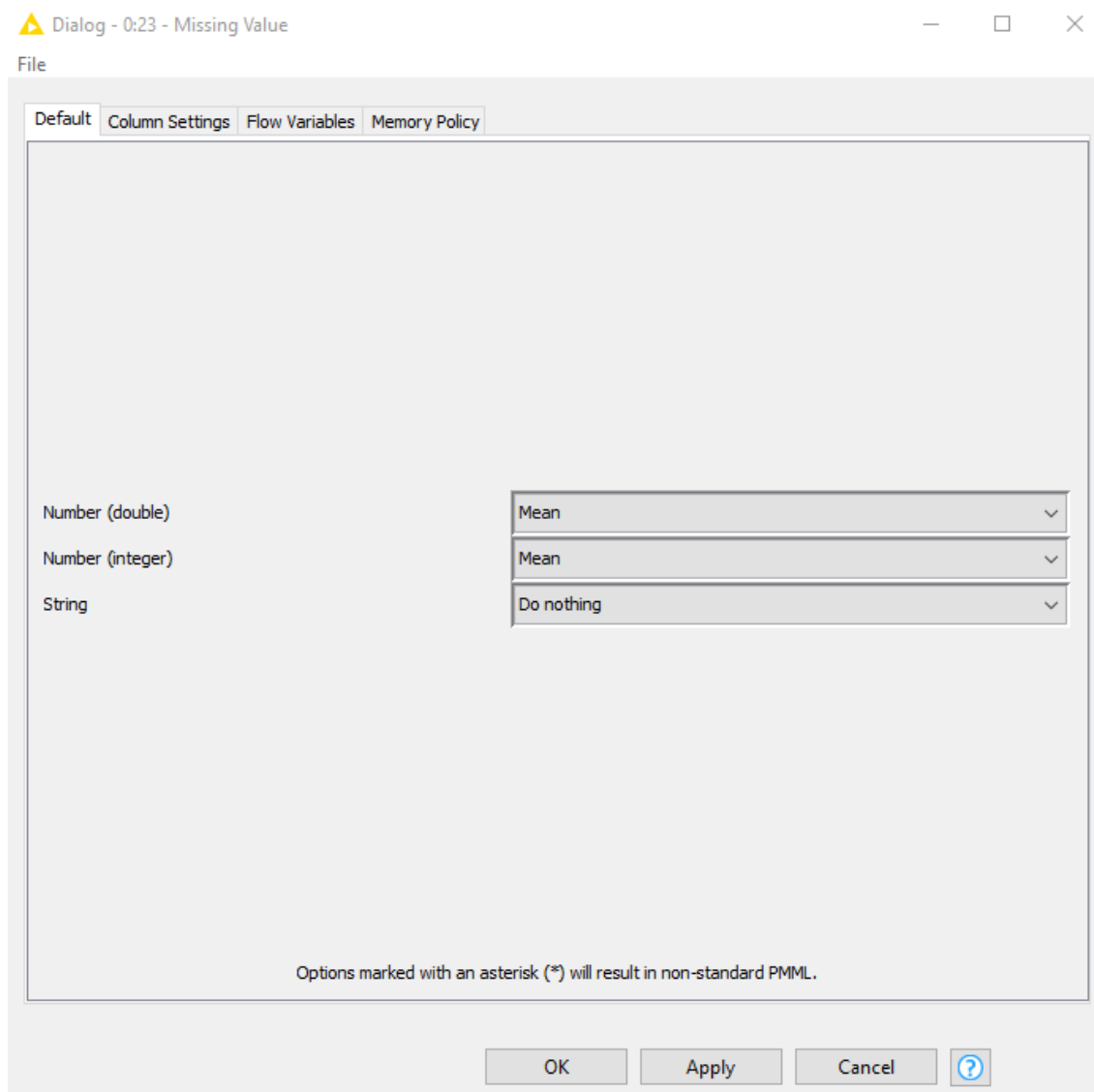
Yiu, T. (2019b). *Understanding Random Forest*. [online] Medium. Elérés helye: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2> [Hozzáférés dátuma: 2021.01.16.].

XI. Mellékletek:

XI.1. A modell beállításai:

A fenti beállítások mellett a véletlen erdő a további alapbeállításokkal rendelkezett, amelyekre mivel nem volt lehetőség átállítani nem tértem ki dolgozatomban részleteiben:

- A bootstrappelt tanuló subset 60%-a volt a teljes tanuló adatnak
- A jellemzők subsetje az összes subset négyzetgyökének felelt meg
- Metszés nélkülvétele
- Nem volt megadva különleges vágási kritérium³²



³² <https://www.youtube.com/watch?v=WdwPOH1C8W0&t=195s>

Dialog - 0:37 - Numeric Outliers

File

Outlier Settings | Group Settings | Flow Variables | Memory Policy

Outlier Selection

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

D script
D clientid

☒ Enforce exclusion

Include

Filter

D debit_type
D balance
D expired_balance
D client_active_from
D product_tv_price_sum
D product_net_price_sum
D product_phone_price_sum
D product_mobilnet_price_sum
D product_other_price_sum

☐ Enforce inclusion

General Settings

Interquartile range multiplier (k) 1.5

Quartile calculation

☐ Use heuristic (memory friendly)

☒ Full data estimate using R_4

☐ Update domain

Outlier Treatment

Apply to All outliers

Treatment option Replace outlier values

Replacement strategy Missing values

OK Apply Cancel ?

Dialog - 0:51 - Partitioning

File

First partition | Flow Variables | Memory Policy

Choose size of first partition

☐ Absolute 100

☒ Relative[%] 93

☐ Take from top

☐ Linear sampling

☒ Draw randomly

☐ Stratified sampling S last_suppl_loy_ext_result

☐ Use random seed 1 617 808 299 7

OK Apply Cancel ?

File

Configuration

Flow Variables

Memory Policy

Nominal column

S net_mig

☒ Use exact sampling

☐ Use approximate sampling

☐ Enable static seed

New Seed

OK

Apply

Cancel



Options Flow Variables Memory Policy

Target Column

Attribute Selection

☐ Use fingerprint attribute

☒ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

D script

D clientid

S created_at

S tv_new

S net_new

S suppl_new

S phone_new

S asset_new

☒ Enforce exclusion

Include

D debit_type

S is_digi_endangered

S is_upc_endangered

S is_flip_endangered

D balance

D expired_balance

D client_active_from

☐ Enforce inclusion

Misc Options

☐ Enable Highlighting (#patterns to store)

☐ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion

☐ Limit number of levels (tree depth)

☒ Minimum node size

Forest Options

Number of models

☐ Use static random seed

Dialog - 0:18 - Random Forest Predictor

File

Prediction Settings Flow Variables Memory Policy

☒ Change prediction column name

Prediction column name

☒ Append overall prediction confidence

☐ Append individual class probabilities

Suffix for probability columns

☒ Use soft voting

OK Apply Cancel ?

Dialog - 0:20 - Scorer

File

Scorer Flow Variables Memory Policy

First Column

Second Column

Sorting of values in tables
Sorting strategy: ☐ Reverse order

Provide scores as flow variables
☐ Use name prefix

Missing values
In case of missing values: ☒ Ignore ☐ Fail

OK Apply Cancel ?

Table 'Default' - Rows: 30 Spec - Columns: 12 Properties Row Variables

Row ID	I TruePo...	I FalsePo...	I TruePo...	I FalsePo...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen...	I Iteration
D#0	14036	182	1792	418	0.971	0.987	0.971	0.908	0.979	:	:	0
I#0	1792	418	14036	182	0.908	0.811	0.908	0.971	0.857	:	:	0
Overall#0	:	:	:	:	:	:	:	:	0.963	0.856	:	0
D#1	14123	202	1681	422	0.971	0.986	0.971	0.992	0.978	:	:	1
I#1	1681	422	14123	202	0.893	0.799	0.893	0.971	0.843	:	:	1
Overall#1	:	:	:	:	:	:	:	:	0.962	0.833	:	1
D#2	14116	167	1746	399	0.973	0.988	0.973	0.913	0.98	:	:	2
I#2	1746	399	14116	167	0.913	0.814	0.913	0.973	0.861	:	:	2
Overall#2	:	:	:	:	:	:	:	:	0.966	0.841	:	2
D#3	14077	186	1763	402	0.972	0.987	0.972	0.905	0.98	:	:	3
I#3	1763	402	14077	186	0.905	0.814	0.905	0.972	0.857	:	:	3
Overall#3	:	:	:	:	:	:	:	:	0.964	0.837	:	3
D#4	14058	173	1757	400	0.972	0.988	0.972	0.91	0.98	:	:	4
I#4	1757	400	14058	173	0.91	0.813	0.91	0.972	0.86	:	:	4
Overall#4	:	:	:	:	:	:	:	:	0.965	0.84	:	4
D#5	14074	185	1741	408	0.97	0.987	0.97	0.904	0.979	:	:	5
I#5	1741	408	14074	185	0.904	0.803	0.904	0.97	0.85	:	:	5
Overall#5	:	:	:	:	:	:	:	:	0.963	0.829	:	5
D#6	14180	142	1730	376	0.974	0.99	0.974	0.924	0.982	:	:	6
I#6	1730	376	14180	142	0.924	0.821	0.924	0.974	0.87	:	:	6
Overall#6	:	:	:	:	:	:	:	:	0.968	0.852	:	6
D#7	14101	209	1722	397	0.973	0.985	0.973	0.902	0.979	:	:	7
I#7	1722	397	14101	209	0.892	0.813	0.892	0.973	0.851	:	:	7
Overall#7	:	:	:	:	:	:	:	:	0.963	0.83	:	7
D#8	14099	209	1740	380	0.974	0.985	0.974	0.903	0.98	:	:	8
I#8	1740	380	14099	209	0.893	0.821	0.893	0.974	0.855	:	:	8
Overall#8	:	:	:	:	:	:	:	:	0.964	0.825	:	8
D#9	14082	192	1774	380	0.974	0.987	0.974	0.902	0.98	:	:	9
I#9	1774	380	14082	192	0.902	0.824	0.902	0.974	0.861	:	:	9
Overall#9	:	:	:	:	:	:	:	:	0.965	0.841	:	9