Adatelemzés és adatvizualizáció R-ben

Logisztikus regresszió az utazási csomagok sikeres adásvételének előrejelzésére

Készítették:

Kürthy Dóra – ILV40L

Hajdu Bálint - VG8M5M

2023. 12. 8.

Tartalomjegyzék

Az adatbázisról röviden

A csoportos beadandónkban használt adatbázisunkat az ajánlott Kaggle oldalról gyűjtöttük be és az ott letöltött adatokkal dolgoztunk.

Az adatbázis kezdetben 20 darab változót tartalmazott, majd mi ezt hosszas gondolkodás után leszűkítettük 12 változóra, amikkel mi a beadandóban meg szeretnénk oldani a feladatokat. Ez a 12 változó megfelel a feladat előírásainak és a megfelelő mennyiségi és minőségi magyarázóváltozókat tartalmazza egy bináris eredményváltozó mellett.

Az adatbázis változói angol nyelven szerepeltek, szóval mi ezt az egyszerűbb értelmezés és hivatkozás kedvéért átalakítottuk magyar nyelvűre. A 12 változó tartalmaz 1 darab bináris eredményváltozót, 6 darab mennyiségi és 5 darab minőségi magyarázóváltozót. Ezeknek a tartalmi jelentései az alábbi felsorolásban olvashatóak:

Az adatbázis tartalma:

- 1 darab bináris eredményváltozó:
  - **Adásvétel:** Megvásárolta-e az utazási csomagot az ügyfél (0=Nem, 1=Igen/Sikeres adásvétel)

- 5 darab minőségi változó:
  - **Kontakt típusa:** Kapcsolatfelvétel kezdeményezője (Iroda, Saját)
  - **Ügyfél neme:** az ügyfél neme (Nő, Férfi)
  - **Preferált színvonal:** az ügyfél által preferált csillagok száma egy hotel esetében (3 csillagos, 4 csillagos, 5 csillagos)
  - **Családi állapot:** ügyfél magánéleti státusza/családi állapota (Egyedülálló, Kapcsolatban, Házas, Elvált)
  - **Pitch értékelése:** ügyfél elégedettsége a pitch-csel (1=Nem tetszett, 2=Kevésbé tetszett, 3=Közömbös, 4=Tetszett, 5=Nagyon tetszett)

- 6 darab mennyiségi változó:
  - **Életkor:** ügyfél életkora (év)
  - **Pitch hossza:** ügynök által tartott pitch hossza az ügyfélnek (perc)
  - **Utazók száma:** ügyféllel utazó emberek száma
  - **Egyeztetések száma:** ügyféllel való egyeztetések száma a pitch után
  - **Utazások száma:** ügyfél átlagos évi utazásszáma
  - **Havi jövedelem:** ügyfél bruttó havi bére (USD)

Az adatbázis azért került kiválasztásra, mert érdekel minket, hogy az általunk választott változók mennyire befolyásolják egy utazás megvalósulását a mintában és a sokaságban is. Az utazás végigkíséri az emberek életét és kíváncsiak vagyunk arra, hogy például a jövedelem, életkor, családi állapot változása és az utazási iroda tevékenységei milyen irányban és hogyan befolyásolják egy utazás megvalósulását.

Outlierszűrés

Az adatbázis sikeres beimportálása, változók átnevezése, hiányzó adatok kiszűrése és a kategória típusú változók faktorrá alakítása után az első dolgunk az outlierszűrés volt a

mennyiségi változóinknál. Az outliersz̋résnél figyelembe vettük, hogy az adatok kiszűrésének maximum 1%-a az elfogadható ezen a téren. Ez a mi adatbázisunknál 41-42 kiszűrhető adatot jelent, mivel 4194 megfigyeléssel rendelkezünk. Az outlierek meglétét dobozábrán vizsgáltunk meg és arra jutottunk, hogy 8 darab megfigyelést kell kiszűrnünk. Az Életkor, Utazók száma és Egyeztetések száma változóinknál nem voltak outlierek, míg a Pitch hosszánál 1, Utazások számánál 4 és Havi jövedelemnél 3 outliert találtunk. Így a kezdeti adatbázisunkat 4194 megfigyelésről 4186-ra csökkentettük, ami azt jelenti, hogy megtartottuk az adatok 99,8%-t. Ez az elfogadható és megengedett tűréshatáron belül van.

Leíró statisztikai elemzés

Az outliersz̋rés után a következő lépés a leíró statisztikai elemzés elvégzése volt a mennyiségi változók esetében. A kapott eredményeket R-ból Excelbe importáltuk és itt láthatóak azok számadatai:

Mennyiségi változók	Gyakoriság	Átlag	Szórás	Medián	Nyesett átlag	Átlagtól vett abszolút távolság	Minimum	Maximum	Terjedelem	Ferdeség	Csúcsosság	Standard hiba
Életkor	4186	37,397	9,242	36	36,936	8,896	18	61	43	0,425	-0,355	0,143
Pitch hossza	4186	15,558	8,236	14	14,581	7,413	5	36	31	0,885	-0,308	0,127
Utazók száma	4186	2,952	0,718	3	2,947	0,000	1	5	4	-0,025	-0,751	0,011
Egyeztetések száma	4186	3,745	1,006	4	3,793	1,483	1	6	5	-0,408	0,616	0,016
Utazások száma	4186	3,290	1,778	3	3,117	1,483	1	8	7	0,870	-0,034	0,027
Havi jövedelem	4186	23359,329	4543,953	22545	22917,132	3378,845	16009	38304	22295	0,790	0,203	70,232

Életkor		Pitch hossza		Utazók száma		Egyeztetések száma		Utazások száma		Havi jövedelem	
Minimum	18,0	Minimum	5,00	Minimum	1,000	Minimum	1,000	Minimum	1,00	Minimum	16009
Első kvartilis	31,0	Első kvartilis	9,00	Első kvartilis	2,000	Első kvartilis	3,000	Első kvartilis	2,00	Első kvartilis	20768
Medián	36,0	Medián	14,00	Medián	3,000	Medián	4,000	Medián	3,00	Medián	22545
Átlag	37,4	Átlag	15,56	Átlag	2,952	Átlag	3,745	Átlag	3,29	Átlag	23359
Harmadik kvartilis	43,0	Harmadik kvartilis	20,00	Harmadik kvartilis	3,000	Harmadik kvartilis	4,000	Harmadik kvartilis	4,00	Harmadik kvartilis	25452
Maximum	61,0	Maximum	36,00	Maximum	5,000	Maximum	6,000	Maximum	8,00	Maximum	38304

1. és 2. ábra: Alap leíró statisztikai elemzések

Az összes számadat értelmezése nagyon sok időt és helyet venne igénybe és nem feltétlen bírna informatív tartalommal az olvasó számára a nyers számadatok egyenkénti értelmezése, ezért csak a legfontosabb, számunkra valóban érdekes és információval bíró adatokat fogjuk megemlíteni.

Az átlag és a medián kapcsolatából, illetve a ferdeségi mutatókból láthatóak a változók eloszlásai. A Pitch hossza, Utazások száma és a Havi jövedelem változók kicsit balra ferdék, ezek a hisztogramon is szemmel láthatóak, ezt igazolja az, hogy az átlag nagyobb, mint a medián és a ferdeségi mutató is pozitív. Az Életkor esetén szintén igazak ezek az állítások, de ez a változó már csak nagyon enyhén balra ferde és a hisztogramon is látható, hogy szinte már normális eloszlást közelít. Az Utazók száma és az Egyeztetések száma változók számadatai egy nagyon enyhe jobbra ferdeségre utalnak, viszont a hisztogramokat megvizsgálva egy szinte normális eloszlást követő diagrammot láthatunk mindkét esetben. Ezt sejteni is lehetett, abból adódóan, hogy az átlag és a medián értékei szinte megegyeznek.

A csúcsossági mutatókat megvizsgálva látható, hogy mindegyik változó értéke 3-nál alacsonyabb, ami azt jelenti, hogy a normális eloszláshoz képest laposabb az eloszlás. Ami még talán érdekesség és említésre méltó, ha a szórás és az átlag kapcsolatát vizsgáljuk, hogy két esetben a szórások a hozzájuk tartozó átlagok felénél nagyobbak. Ez a két eset a Pitch hossza és az Utazások száma változók. Ezekben az esetekben a legnagyobbak a szórások a mennyiségi változók tekintetében, mert itt a relatív szórás 50%-nál nagyobb értéket vesz fel mindkét esetben.

Amit még érdekesnek találtunk, az az, hogy a mintában az ügyfelek 75%-a 43 éves vagy annál fiatalabb, ami egy elég fiatalos adatbázist szolgáltat számunkra. Látható, hogy egy utazási csomag ismertetése átlagosan 15 és fél percet vesz igénybe az utazási iroda értékesítőjének és

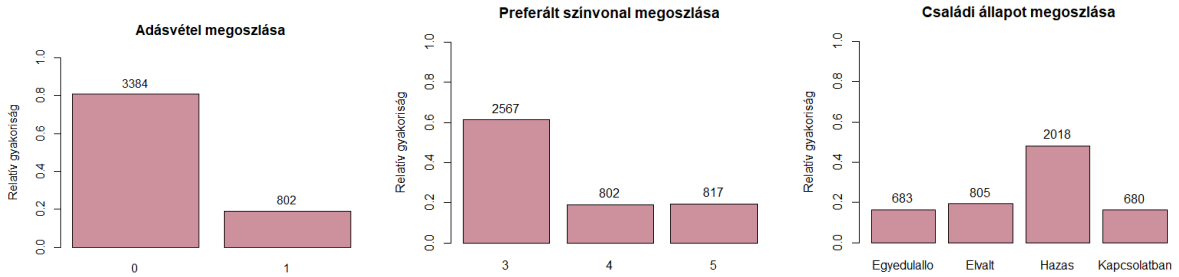
átlagosan 4 egyeztetés szokott lezajlani a munkatársak és az ügyfelek között a pitch ismertetése után. Az ügyfelek fele legalább 3 utazásra megy el egy évben, amely szám elég nagynak mondható, hogy ha nyaralásról van szó és nem üzleti útról. Ebből talán következtethetünk arra is, hogy a mintánkban szereplő ügyfelek a tehetősebb kategóriába tartoznak a vagyoni elemeket illetően. Ezt igazolja az is, hogy az átlagos Havi jövedelem 23 359 dollár, ami magyar forintra átszámolva jelenlegi árfolyammal körülbelül 8,3 millió forint. Ez az összeg nem a kilógó értékek miatt ilyen magas, mert már a minimum Havi jövedelem is 16 009 dollár, ami nagyjából 5,7 millió forintnak felel meg.

Ezen statisztikai adatok alapján tehát látható, hogy a mintában szereplő megfigyelések inkább a fiatalabb és vagyonosabb személyeket tartalmazza, akik egy évben többször is hajlandóak elutazni üdülés vagy akár munka szempontjából. Ennek következtében nem valósul meg teljesen a FAE mintavétel, mert az utazási iroda célzottan magasabb jövedelmű célcsoportot keresett meg, hogy növelje esélyeit és bevételeit az értékesítés során. Ezzel egy torzított mintacsoport áll rendelkezésünkre az adatbázisban, de a nagy elemszám kompenzálja ezt számunkra.

Adatvizualizációs elemzés - Előfeltevéseink

Egyváltozós adatvizualizáció:

Az adatvizualizációs részt a változók eloszlásainak vizsgálatával kezdtük. Itt az órákon tanultak alapján a legjobb módszert alkalmaztuk, amely vizualizáció az oszlopdiagram megjelenítése lesz, mivel ez preferált a tudományos munkákban is. Az emberi szem jobban tud lineáris mértéket kezelni, mint területet a kördiagram esetében. Az oszlopdiagram szempontjából a magasság csak egydimenziós, a kördiaram viszont 2 dimenziós. Ezen gondolatmenet után tehát a minőségi változókat oszlopdiagramon, a mennyiségi változókat pedig hisztogramon ábrázoltuk. Az egyváltozós vizualizációkból az R-ben 6-6 darab készült mindkettő változótípus esetében, ezért mi ezek közül csak 3-3-at fogunk a dokumentumban megjeleníteni a helytakarékosság érdekében.

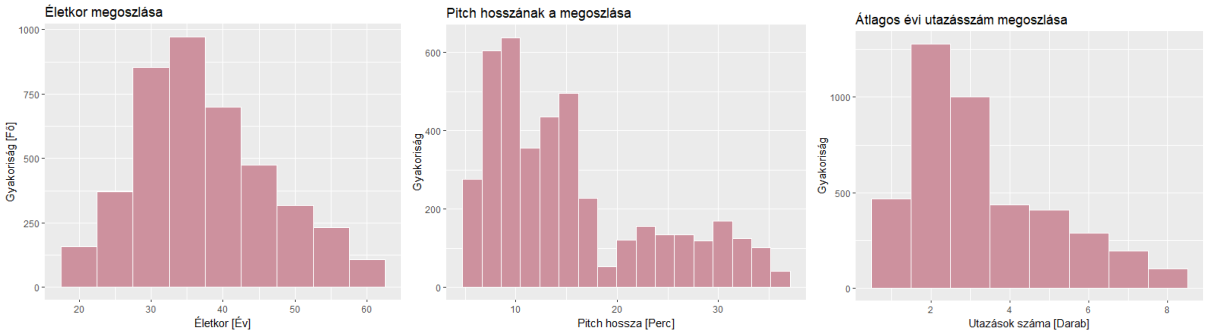


3., 4. és 5. ábra: Megoszlások oszlopdiagrammon

Az oszlopdiagramokat Base R vizualizációval valósítottuk meg ebben az esetben, mivel ennek a használata és az output eredménye preferáltabb számunkra. Az adásvétel megoszlásánál látható, hogy jóval nagyobb arányban van az a kimenetel, hogy az ügyfél elutasítja az utazási csomagot, nagyjából minden 5 potenciális utazási csomag adásvételéből 1 valósul meg. A preferált színvonal megoszlásán látható, hogy a legjobban a 3 csillagos hoteleket preferálják az ügyfelek. Ez egy kicsit meglepett minket, mivel amint már említést tettünk róla, a minta a tehetősebb ügyfeleket tartalmazza és ebből arra számítottunk, hogy az 5 csillagos

hotelek esetében lesz ez a kimagasló arány. A családi állapotnál látható, hogy nagyjából az ügyfelek fele házas és a többi kategóriában szinte egyenlően oszlanak meg az adatok.

A mennyiségi változók esetében a hisztogramokat az R-ben ggplot-tal ábrázoltuk, mert ezen vizualizáció terén ezt preferáljuk jobban. Az ábrákról nem tennénk különösebb megjegyzéseket, mivel már minden szükséges információt leírtunk ehhez a részhez a leíró statisztikai elemzés című bekezdésben. Az ott tett állításainkat vizuálisan is látni és ellenőrizni lehet.



6., 7. és 8. ábra: Megoszlások hisztogramon

Asszociációs kapcsolatok

Minden kategória típusú magyarázó változó kapcsolatát megvizsgáltuk kereszttáblában az eredményváltozóval (ami szintén kategória típusú, ezért asszociációs a kapcsolat).

A Preferált színvonal esetében azt láttuk, hogy a 3-4 csillagos szállások közel azonos arányban kerültek megvételre (16,5% és 19,8%), az 5 csillagosok viszont jelentősen nagyobb arányban (26,9%). Mivel a 3 csillag a referencia, ezért ennél a változónál pozitív bétát várunk, hiszen az ettől való eltérés esetén növekedik az adásvétel aránya.

A Pitch értékelése változó esetén nem konzisztensen nő az adásvétel aránya a pitch pontszámának növekedésével: az jól látszódott, hogy az 1-2-3-5 között növekvő tendencia van, de a 4-es egy kilógó érték, itt kisebb volt az adásvétel aránya, mint a 3-asnál. Ezért megnéztük, emögött milyen hatás (volumen vagy összetétel) állhat: 3-as értékelést adnak legtöbben és 4-est legkevesebben, ami azon emberi magatartásból fakadhat, hogy valamit vagy közepesre vagy nagyon jóra értékelünk, ritka, hogy enyhe különbségekre hajlunk (vagy közömbös vagy nagyon tetszett, ritka, hogy „tetszett"). Ezen változóra tehát enyhén pozitív bétát várunk, illetve, hogy nem feltétlen lesz szignifikáns.

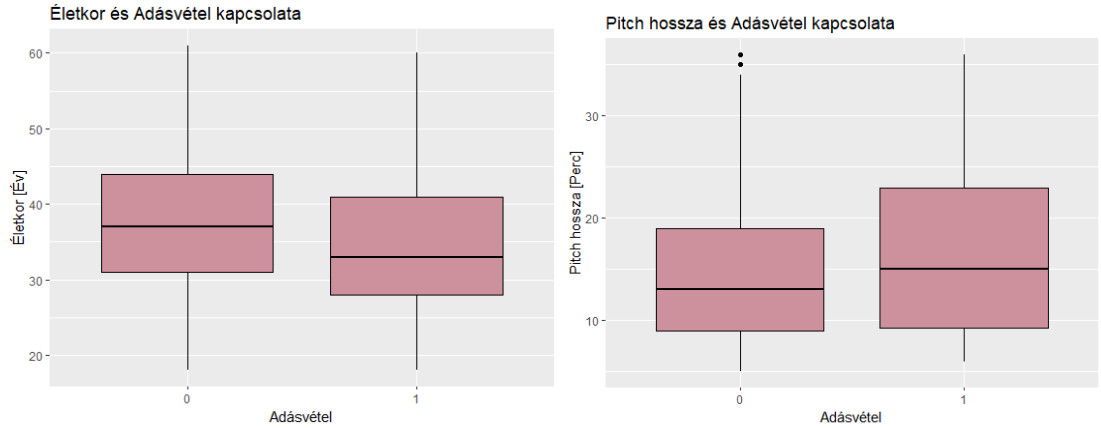
Az Ügyfél neme változó esetén azt láttuk, hogy közel azonos arányban vásároltak, illetve nem vásároltak a nők és a férfiak (mintaarány 19,2% környékén mindkét esetben) így erre a változóra azt várjuk, hogy nem lesz szignifikáns.

A Családi állapot változó esetén az adásvétel legmagasabb aránya az Egyedülálló kategóriában mutatkozott (0,3566), és mivel ez a referencia kategóriánk is, az ettől való eltérés csökkenti az adásvétel sikerének arányát, így negatív bétát várunk a változóra az alábbi sorrendben: Egyedülálló > Kapcsolatban > Elvált > Házas. Emögött az életkor előrehaladtát tudjuk elképzelni magyarázatként, a házasok és elváltak általában idősebbek és már jobban leköti őket a munkájuk, családjuk, míg az egyedülállók és kapcsolatban lévők jellemzően fiatalabbak és kötetlenebbek utazás terén.

A Kontakt típusa változó esetén láthatjuk, hogy az iroda általi felkeresés nagyobb arányban lett sikeres kimenetelű (0,226), mintha valaki önként kereste volna fel az irodát (0,178), mivel az Iroda a referencia kategória, negatív bétát várunk a változóra. Ennek lehetséges oka lehet, hogyha valaki proaktívan keresi fel az irodát, akkor jellemzően van egy körvonalazott elképzelése az útról, és ha ezt nem találja meg, akkor odébb áll, míg, ha véletlenszerűen kap az irodától egy ajánlatot, nincsenek előzetes elvárásai, lehet megtetszik neki és elviszi.

Kétváltozós adatvizualizáció: Vegyes kapcsolatok

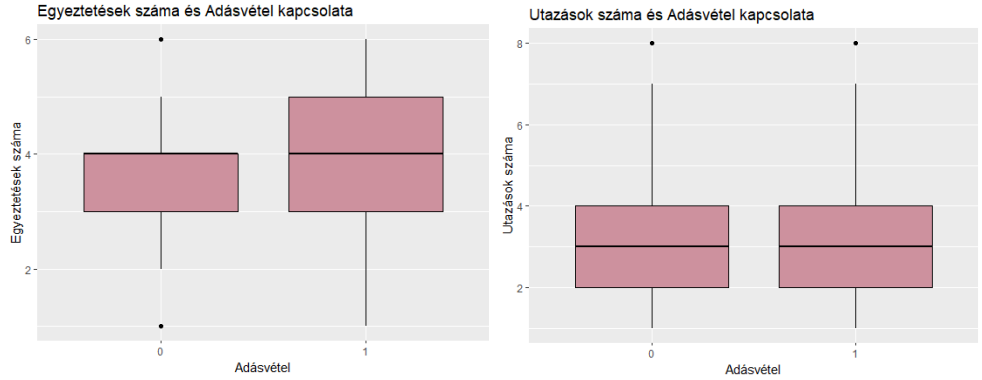
Az eredményváltozó és a mennyiségi magyarázó változók kapcsolatára páronkénti dobozábrákat vizsgáltunk.



9. és 10. ábra: Vegyes kapcsolatok dobozábrán I.

Az Életkor változó esetén a sikeres adásvétel Életkor mediánja alacsonyabb, mint a nem sikeres kimenetelnél, így azt állapíthatjuk meg, hogy inkább a fiatalabbak vásárolták meg végül a csomagot, tehát az adásvétel Oddsát csökkenti a negatív bétát várunk a változóra. Ez összecseng tehát a korábbiakban Családi állapot változóra tett feltevésünkkel, miszerint a fiatalabbak kevésbé kötöttek, akár munka akár családi állapot okán, így inkább hajlandóak elfogadni az utazási ajánlatot.

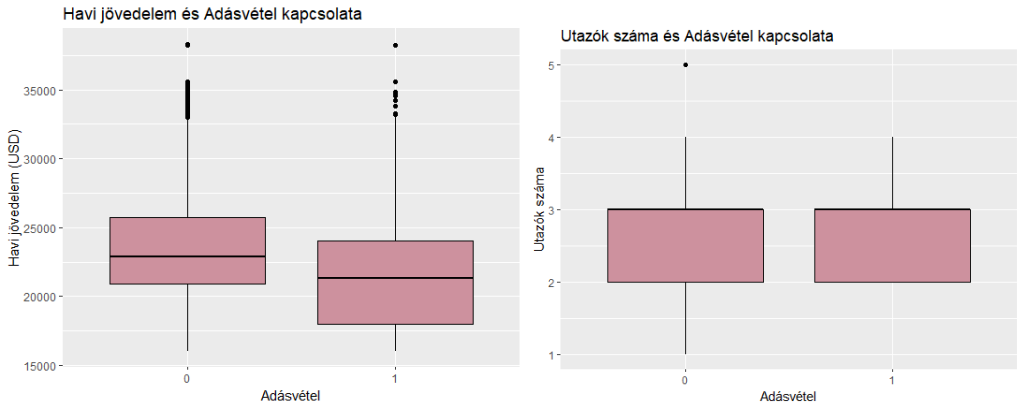
A Pitch hossza változónál a sikeres pitch-csek mediánja némileg magasabb a nem sikeresekénél, tehát a hosszabb pitch-csek inkább eredményeztek sikeres adásvételt, így enyhén pozitív bétát várunk.



11. és 12. ábra: Vegyes kapcsolatok dobozábrán II.

Az Egyeztetések száma változónál volt, aki 1 egyeztetés után rögtön elfogadta az ajánlatot és volt, aki a 6. után lépett vissza, a medián mindkét kimenetel esetén 4 egyeztetés, így nem várunk jelentős bétát, a dobozábra alapján enyhén pozitívat.

Az Utazások száma változó nem mutat eloszlásbeli különbséget a sikeres és sikertelen kimenet között, mindkét esetben évi 3 út a medián, nem várjuk szignifikánsra ezt a változót.



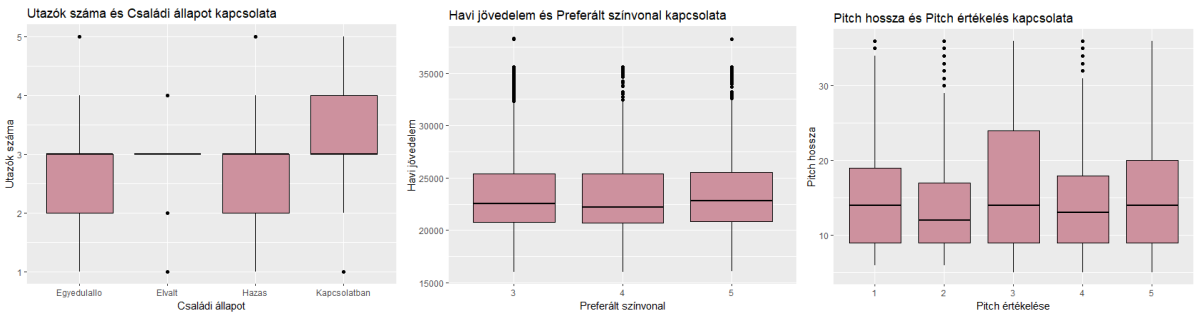
13. és 14. ábra: Vegyes kapcsolatok dobozábrán III.

A Havi jövedelem változó esetén a vevők között kisebb a mediánbér, mint a nem-vevők között, így enyhén negatív bétát várunk a változóra. Ez elsőre érdekes lehet, hiszen azt gondolnánk, akinek van pénze, az többet is tud költeni, ám itt mégis a kevesebb jövedelműek tűnnek inkább elfogadónak a kiadásra. Ez összefügghet azzal, hogy láttuk, a fiatalabbak körében volt nagyobb a sikeres adásvétel, feltehetőleg a fiatalabbaknak kevesebb is a jövedelme, ezért lehet, hogy a kevesebb jövedelműek jelennek meg itt is nagyobb arányban, tehát Életkoruk áll a háttérben. Erre kiváló lenne útelemzést végrehajtani, de ez most nem dolgozatunk témája. Helyette megnézzük majd korrelációban mit jelent ez a két változó között. Másrészt ez az információ fakadhat abból is, hogy akinek több a pénze, tudatosabban bánik vele, nem költi el olyan könnyedén, lehet emiatt is van neki sok.

Az Utazók számánál nem láthatunk semmi különbséget, nem várjuk szignifikánsra ezt a változót.

<u>Kétváltozós adatvizualizáció: Multikollinearitás esetleges előrejelzése</u>

Általunk intuitíve logikusnak vélt összefüggéseket néztünk meg, hogy felfedezzünk esetleges kapcsolatokat a magyarázó változók között a „by” függvény segítségével csoportosításban. A kapcsolatok dobozábrái itt láthatóak:



15., 16. és 17. ábra: Multikollinearitás előrejelzése dobozábrán

A Családi állapot és az együtt utazók száma esetén azt vártuk, hogy aki házas, az jellemzően családdal, azaz többen utazik, ám nem láttunk lényegi különbséget a csoportok

között, mindenhol 3 fő együtt utazó volt a medián. Ennek oka lehet, hogy az egyedülállók jellemzően barátokkal, azaz szintén többen utaznak.

Második feltevésünk az volt, hogy magasabb havi jövedelemhez magasabb színvonal preferálása fog társulni, ám ez sem lett igaz, úgy tűnik a jövedelem szintje nem befolyásolta a minőséget illető választást, volt, aki gazdag mégis spórolna, és volt, aki nem annyira tehetős mégis 5 csillagosba szeretne menni, ez szintén az emberi viselkedés formájából adódik.

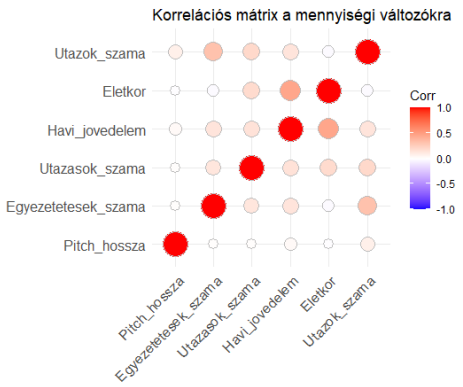
Harmadik feltevésünk pedig, hogy a Pitch értékelése függ a Pitch hosszától, hiszen minél hosszabb, annál meggyőzőbb tud lenni az ügynök, ám ez sem bizonyult igaznak.

Többváltozós adatvizualizáció

A ggplot2 segítségével korrelációs mátrixot készítettünk összes mennyiségi változónk bevonásával, melynek eredménye az alábbi:

Az Életkor és a Havi jövedelem korrelál legjobban egymással, körülbelül 0,4-es értékben pozitív irányban, amire számítottunk is. Így már egyértelmű, hogy az Életkor előrehaladtával nő a Havi jövedelem, emiatt mindkét változó paraméterére negatív értéket várhatunk a regresszióban.

A többi változó legfeljebb 0,2-es mértékben korrelál egymással, így nem számítunk magas VIF értékekre modellünkben.



18. ábra: Korrelációs mátrix

Regressziós modell

Mielőtt a regressziós modellről írnánk, említést tennénk arról, hogy a minőségi változók referenciakategóriáit még a változók átalakításánál megvizsgáltuk és ahol szükséges volt, ott megváltoztattuk a viszonyítási alapot. Ez a változtatás csak a Családi állapot esetében történt meg, itt az Egyedülálló kategóriát gondoltuk úgy, hogy jó viszonyítási alapot ad majd nekünk. A többi változónál megfeleltek az alap referenciák, ami a Kontakt típusa esetében az Iroda, az Ügyfél neme esetében pedig a Nő kategóriák voltak.

Ezek után következhetett a regressziós modell futtatása egy olyan modell esetében, amely tartalmazza az összes magyarázóváltozót. Ezt mi modell_alap-nak neveztük el az R-ben. Ez a modell első ránézésre elég jónak nézett ki nekünk, a p-értékeket vizsgálva nagyon sok szignifikáns változót kaptunk, amely azt jelenti, hogy ezek a változók nem csak a mintában, hanem a sokaságban is szignifikáns változók és jól magyarázzák az eredményváltozónkat. Ez azt is jelenti számunkra, hogy kevés magyarázóváltozót kell majd csak kiszűrnünk a modellszűkítésnél. Hogy pontosan melyek ezek a változók, azokat majd tisztán a sorbarendezett

p-értékek alapján láthatjuk. Itt láthatóak a lefutatott regresszió paraméterei az összes magyarázóváltozóval:

	Estimate	Std. Error	z value	Pr(>|z|)
Konstans	-0,189	0,331	-0,569	0,569
Életkor	-0,026	0,005	-4,755	0,000
Kontakt típusaSaját	-0,290	0,090	-3,235	0,001
Pitch hossza	0,031	0,005	6,342	0,000
Ügyfél neme: Férfi	0,159	0,086	1,840	0,066
Utazók száma	-0,079	0,065	-1,220	0,223
Egyeztetések száma	0,357	0,048	7,517	0,000
Preferált színvonal: 4	0,215	0,110	1,957	0,050
Preferált színvonal: 5	0,752	0,102	7,344	0,000
Családi állapot: Elvált	-1,195	0,142	-8,433	0,000
Családi állapot: Házas	-1,167	0,111	-10,504	0,000
Családi állapot: Kapcsolatban	-0,501	0,131	-3,812	0,000
Utazások száma	0,065	0,024	2,654	0,008
Pitch értékelése: 2	0,068	0,169	0,400	0,689
Pitch értékelése: 3	0,371	0,126	2,939	0,003
Pitch értékelése: 4	0,187	0,142	1,313	0,189
Pitch értékelése: 5	0,469	0,137	3,427	0,001
Havi jövedelem	0,000	0,000	-6,135	0,000

19. ábra: Alapmodell koefficens táblázata

A táblázatban az értékeket 3 tizedesjegyig jelenítettük csak meg a könnyebb láthatóság érdekében és pirossal kiemeltük azokat a változókat/kategóriákat, amelyek nem minden szokásos szignifikanciaszinten szignifikánsak. A konstanst nem vettük ebben az esetben figyelembe. A Férfi ügyfelek kategória 10%-on még igen, de 5%-os alfán már nem szignifikáns változó. Az utazók száma változó és a 2-es és 3-as Pitch értékelés kategóriák pedig egyik szokásos szignifikanciaszint mellet sem szignifikánsak.

A változók bétáira rátekintve leellenőrizhetőek a vizualizációnál tett előfeltevéseink. Minden esetben valóban jól előrejeleztünk egy változót azon a téren, hogy szignifikáns lesz-e vagy éppen, hogy milyen irányú hatással van az egységnyi változása az adásvételre.

Modellszelekció

A modellszelekciót ezen eredmények után kézi és gépi módon is elvégeztük és utána összehasonlítottuk az eredményeket mindhárom modell esetében. A modellszűkítést a kézi, azaz manuális módszerrel kezdtük. Itt a p-értékek alapján egy fontossági sorrendet állítottunk, amelynek eredménye itt látható:

	Estimate	Std. Error	z value	Pr(>|z|)
Családi állapot: Házas	-1,167	0,111	-10,504	0,000
Családi állapot: Elvált	-1,195	0,142	-8,433	0,000
Egyeztetések száma	0,357	0,048	7,517	0,000
Preferált színvonal: 5	0,752	0,102	7,344	0,000
Pitch hossza	0,031	0,005	6,342	0,000
Havi jövedelem	0,000	0,000	-6,135	0,000
Életkor	-0,026	0,005	-4,755	0,000
Családi állapot: Kapcsolatban	-0,501	0,131	-3,812	0,000
Pitch értékelése: 5	0,469	0,137	3,427	0,001
Kontakt típusaSaját	-0,290	0,090	-3,235	0,001
Pitch értékelése: 3	0,371	0,126	2,939	0,003
Utazások száma	0,065	0,024	2,654	0,008
Preferált színvonal: 4	0,215	0,110	1,957	0,050
Ügyfél neme: Férfi	0,159	0,086	1,840	0,066
Pitch értékelése: 4	0,187	0,142	1,313	0,189
Utazók száma	-0,079	0,065	-1,220	0,223
Konstans	-0,189	0,331	-0,569	0,569
Pitch értékelése: 2	0,068	0,169	0,400	0,689

20. ábra: Változók fontossági sorrendje p-érték alapján

A jelölések megegyeznek az előző bekezdésben használatos jelölésekkel. Itt látható, hogy az említett nem szignifikáns változók/kategóriák a sorrendben az utolsó helyeket foglalják

el. A konstanst szintén nem kell itt sem figyelembe venni. Az eredmények alapján mi 3 változót szűrnénk ki, amelyek a következőek: Pitch értékelése, Utazók száma, Ügyfél neme. A Pitch értékelése változónak 2 kategóriája ugyan szignifikáns, de a 2-es és 4-es értékelések magas p-értéke miatt mi azt a döntést hoztuk, hogy ki kell szedni ezt a változót is. A Férfi Ügyfél kategória kiszedése mellett mi azért döntöttünk, mert a modellben mi csak a minden szokásos alfán szignifikáns változókat szeretnénk benntartani és ez itt nem valósul meg. Ennek következtében létrehoztunk egy új modellt, amely az említett változókat nem tartalmazza. Ezt mi modell_szukitett_kezi-nek neveztünk el az R-ben.

A következő lépés a gépi modellszelekció elvégzése volt, amelyet Stepwise módszerrel hajtottunk végre. Eredményül azt kaptuk, hogy a gép 4 változót dobna ki, melyek sorrendben a következőek: Pitch értékelése, Utazók száma, Ügyfél neme és Utazások száma. Az első 3 változót mi is eliminátuk, de a gép még egy negyediket is kidobott a modellből, amely eredetileg egy minden szokásos alfán szignifikáns változó. Ezen eredmények alapján létrehoztunk egy újabb modellt, amelyet modell_szukitett_gepi-nek neveztünk el az R-ben.

Az eredmények és az új modellek létrehozása után összehasonlítottuk az eredményeket az információs kritériumok eredményei alapján és ezek számadatai itt láthatóak:

Modellek	Szabadságfokok	AIC	BIC
modell_alap	18	3668,247	3782,358
modell_szukitett_kezi	12	3678,189	3754,263
modell_szukitett_gepi	11	3683,300	3753,034

21. ábra: Modellek összehasonlítása információs kritériumok alapján

Az AIC értékelése alapján a sorrend: modell_alap > modell_szukitett_kezi > modell_szukitett_gepi. A BIC értékelése alapján a sorrend pedig megfordul. Az AIC kevésbé bünteti a plusz magyarázóváltozók számát, ezért lehetséges, hogy az alapmodell a legjobb ezen kritérium alapján. A BIC alapján viszont már a gépileg szűkített modell lesz a preferált. Mi ezek ellenére is az általunk manuálisan szűkített modellel szeretnénk dolgozni, mivel ahogyan már említettük, a gépi elimináció során egy minden szokásos alfán szignifikáns változót vesztettünk el az Utazások száma változó esetében. Mi ezt a folytonos változót meg szeretnénk tartani és nem szeretnénk kihagyni a modellünkből. Ha ezt a döntést vizsgáljuk, akkor AIC alapján a manuálisan szelektált modell preferált a gépivel szemben, míg a BIC alapján csak minimális az eltérés. Ezen minimális eltérésből adódóan egy könnyed egyszerűsítéssel az mondható el, hogy mindkettő modell egyenlően preferált és nem érzékelhető különbség a további elemzési eredményeinkben, ha a manuális modellt választjuk. Ezek után megvizsgáltuk a McFadden féle pszeudo R-négyzeteket is mindhárom modell esetében. Amit kiemelnénk az az, hogy ha a manuális modellt választjuk, akkor csak 0,54%-ponttal csökken az R-négyzet az alapmodellhez képest, ami nagyon kevésnek mondható és ez is azt igazolja, hogy jól döntünk, ha a manuális modellt választjuk. A mutató értelmezésére a következő bekezdésben bővebben is szót ejtünk.

Ezek után az R-ben az egyszerűsítés és könnyebb értelmezhetőség érdekében létrehoztuk a végső modellt, amelyet modell_vegleges-nek neveztünk el és ez teljesen egyenértékű a manuális modellünkkel, amely a modell_szukitett_kezi névre hallgat. A végső modell eredményei itt láthatóak:

	Estimate	Std. Error	z value	Pr(>|z|)
Konstans	0,050	0,298	0,168	0,867
Életkor	-0,026	0,005	-4,769	0,000
Kontakt típusaSaját	-0,308	0,089	-3,462	0,001
Pitch hossza	0,032	0,005	6,557	0,000
Preferált színvonal: 4	0,206	0,109	1,884	0,060
Preferált színvonal: 5	0,737	0,102	7,250	0,000
Egyeztetések száma	0,338	0,045	7,472	0,000
Családi állapot: Elvált	-1,217	0,138	-8,804	0,000
Családi állapot: Házas	-1,185	0,110	-10,799	0,000
Családi állapot: Kapcsolatban	-0,540	0,130	-4,158	0,000
Utazások száma	0,065	0,024	2,683	0,007
Havi jövedelem	0,000	0,000	-6,312	0,000

22. ábra: Végső modell koefficens táblázata

Globális Khi-négyzet próba

A legjobbnak tartott, azaz végső modellelünkön elvégeztük a GOF, azaz Goodness of Fit tesztet. Itt a H_0, azaz nullhipotézisünk az, hogy $B_1 = B_2 = B_3 \ldots B_k == 0$, azaz a sokaságban összeomlik a minta magyarázóereje. A H_1 ezzel szemben azt állítja, hogy van legalább egy B_j, ami nem 0, azaz szignifikáns. Ebben az esetben a modell kiterjeszthető a sokaságra, vagyis a valóságra. Ennek a hipotézisnek az eredményét Khi-négyzet próba alapján tudjuk megkapni. A Khi-négyzet próbafüggvény értéke a nulldeviancia és a reziduális deviancia (saját modell) különbsége. A mi esetünkben a nulldeviancia 4089,9, ez azt jelenti, hogy ez az üres modellnek a devianciája. Pontosabban a tökéletes modelltől való eltérés. A tökéletes modell devianciája 0. A reziduális deviancia a mi esetünkben 3654,2 és ez is a tökéletes modell 0 devianciájától való távolságot mutatja. Ezek után megkaptuk, hogy a próbafüggvényünk értéke 435,7. A szabadságfokok száma 8, mivel ennyi magyarázóváltozót tartalmaz a modellünk. Ez egy jobboldali próba, tehát a feléesési valószínűséget keressük. A próbafüggvény kiszámítása után megkaptuk, hogy a p-érték 0. Ennek nagyon örültünk, mert ez azt jelenti, hogy a H_0-t elutasítjuk és az aktuális modell mellett döntünk. Ez továbbá azt is jelenti, hogy a modell releváns a sokaságban is.

Magyarázóerő kifejtése

Pszeudo R négyzet kézi kiszámítása

Végső modellünkre a beépített függvénnyel 10,65%-os pszeudo R négyzet értéket kaptunk, amit érdekesnek tartottunk megvizsgálni kicsit mélyebben, miből is tevődik össze ez az éppen már közepesnek mondható, de a gyengéhez közelebb álló magyarázóerő.

A Khi-négyzet próba elvégzésekor már volt szó az üres modell és az aktuális modell devianciájának eltéréséről, hiszen ez volt a próbafüggvény értéke. Itt ezeket az értékeket használjuk a magyarázóerő kézi kiszámításához.

pszeudo R négyzet = (üres modell dev. – aktuális modell dev.) / üres modell dev.

Ez az úgynevezett Likelihood Ratio alapján működik, melynek lényege, hogy a tökéletes modelltől való távolságot veszi figyelembe, ami valójában a 0-tól vett távolság, hiszen egy tökéletes modell minden megfigyelésre 100%-os valószínűséggel mond helyes értéket, azaz minden tagja ln1 = 0.

Az alábbi eredményt kaptuk: (4089.9 - 3654.2) / 4089.9 = 10,65%

Az eredmény nem meglepő módon ugyanaz, mint gépi kiszámítással, ám az értelme sokkal inkább látható: *végleges modellünk 8 magyarázó változó segítségével az üres modelltől (4089.9) a szaturált modellig (0) való távolság 10,65%-át tette meg.* Ez összecseng a khi-négyzet próba eredményével, miszerint szignifikáns modellünk eltávolodása az üres modelltől.

VIF – mutató

Másik fontos mutatónk a multikollinearitást mérő VIF – mutató, melyhez az adatvizualizáció egy része kapcsolódott. A magyarázó változók közötti kapcsolatok vizsgálatánál végső konklúziónk az volt, hogy nem számítunk jelentős multikollinearitásra, ám ezt a modellépítés után mindenképp leteszteljük.

VIF – mutatónk korrigált formában jelenik meg (feltehetőleg a sok kategória típusú változó okán), és minden érték 2 alatt van, ami roppant alacsony multikollinearitást jelent, modellünk ezen a téren jól teljesít.

Regressziós egyenlet

Esélyhányadosunk (Odds) = P (sikeres adásvétel) / 1-P (sikertelen), így a pozitív előjelű Béta paraméterek növelni fogják a sikeres adásvétel Odds-át, a negatívak pedig csökkenteni azt. Az esélyhányados logaritmusa, azaz a logit, felírható a magyarázó változók lineáris összefüggéseként (ln Odds = ln (P/1-P) = alfa + béta * X), tehát magát az esélyhányadost (becsült y) a magyarázó változók exponenciális összefüggéseként lehet értelmezni (P = e ^ logit / (1 + e ^ logit), ahol logit = alfa + béta * X), ennek okán a végleges modellben kapott Béta paramétereket e – adra emelve tudjuk értelmezni, melyet most 3 tizedesre kerekítve szemléltetünk, kivéve a jövedelem esetén, ahol több szükséges:

- Odds adásvétel (y kalap) =

  *1,052 + 0,975 * Életkor + 0,735 * Kontakt_típusa:Saját + 1,033 * Pitch hossza*

  *+ 1,228 * Preferált színvonal:4 + 2,090 * Preferált színvonal:5*

  *+ 1,402 * Egyeztetések száma + 0,296 * Családi állapot:Elvált*

  *+ 0,306 * Családi állapot:Házas+ 0,583 * Családi állapot:Kapcsolatban*

  *+ 1,067 * Utazások száma + 0,99992 * Havi jövedelem*

Paraméterek értelmezése

Minden változó paraméterének értelmezésekor feltételezzük, hogy a többi változó változatlan, azaz ceteris paribus tekintjük a Bétákat.

Értelmezések:

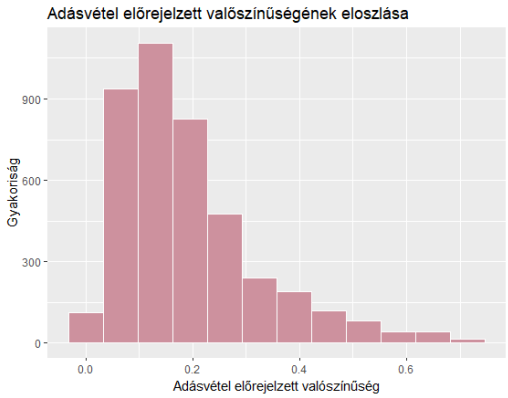
- A **konstans** azt mutatja számunkra, hogy egy minden változó tekintetében 0 tulajdonságú ügyfél esetén az adásvétel Odds 1,052 lenne, ez azonban nem az értelmezési tartományunk része az Életkor változó miatt, hiszen 0 éves személy nincs (egyébként a többi változó esetén még lenne is értelme az X=0 helynek, hiszen a mennyiségik közül 0 egyeztetés a pitch után lehetséges (valaki rögtön elfogadta) és 0 évi utazás is megeshet (eddig nem utazott, most kezdene el utazgatni), a kategória típusúaknál pedig a referencia kategóriát jelölné a 0-ás kimenet, ami szintén értelmes,

esetleg még a havi jövedelem 0 is elképzelhető, ha más fizeti ki az utat számára, de az ő nevére kérik mégis a számlát). Így tehát az 1,052 egy illeszkedést segítő paraméter csupán.

- **Életkor:** Ha egy ügyfél egy évvel idősebb, várhatóan 2,5%-kal csökken a sikeres adásvétel Oddsa.
- **Kontakt típus: Saját:** Ha valaki saját keresés alapján kerül kapcsolatba az utazási irodával, akkor várhatóan 26,5%-kal kisebb az adásvétel Oddsa, ahhoz képest, mintha az utazási iroda kontaktálná.
- **Pitch hossza:** Ha az utazási iroda által tartott pitch 1 perccel hosszabb, akkor az ügyfél adásvétel Oddsa várhatóan 3,3%-kal több.
- **Preferált színvonal: 4 csillag**: Ha az ügyfél által preferált szálláshely 4 csillagos, akkor az ügyfél adásvétel Oddsa várhatóan 22,8%-kal több, mintha 3 csillagos lenne.
- **Preferált színvonal: 5 csillag:** Ha az ügyfél által preferált szálláshely 5 csillagos, akkor az ügyfél adásvétel Oddsa várhatóan 109%-kal több, mintha 3 csillagos lenne.
- **Egyeztetések száma:** Ha az ügyfél és az iroda között 1-gyel több egyeztetésre kerül sor, akkor az ügyfél adásvétel Oddsa várhatóan 40,2%-kal több.
- **Családi állapot: Elvált:** Ha egy ügyfél Elvált, akkor az adásvétel Oddsa várhatóan 70,4%-kal kisebb, mintha Egyedülálló lenne.
- **Családi állapot: Házas:** Ha egy ügyfél Házas, akkor az adásvétel Oddsa várhatóan 69,4%-kal kisebb, mintha Egyedülálló lenne.
- **Családi állapot: Kapcsolatban:** Ha egy ügyfél Kapcsolatban él, akkor az adásvétel Oddsa várhatóan 41,7%-kal kisebb, mintha Egyedülálló lenne.
- **Utazások száma:** Ha egy ügyfél évente átlagosan 1-gyel többször utazik, akkor az adásvétel Oddsa várhatóan 6,7%-kal több.
- **Havi jövedelem:** Ha egy ügyfél havi bruttó jövedelme 1 dollárral több, akkor az adásvétel Oddsa várhatóan 0,008%-kal kisebb. Ez érthetőbben: Ha egy ügyfél havi bruttó jövedelme 1000 dollárral több, akkor az adásvétel Oddsa várhatóan 8%-kal kisebb.

Klasszifikációs mátrix elkészítése és értelmezése

Első lépésben a predict.glm függvény segítségével data frame-ünkhöz hozzáadtunk 2 új oszlopot: Logit és Probability, melyek minden egyes megfigyeléshez azok adottságait a végleges modellbe helyettesítve megadják ezen értékeket. Y = 1: a sikeres adásvételt tekintjük pozitív osztálynak, erre vonatkozóan szeretnénk mutatóinkat meghatározni. Cél, hogy modellünk a sikeres adásvételt jelezze előre pontosabban, hogy az utazási iroda jól tudjon tervezni kapacitással, inkább több lehetőséget tudjanak kínálni ügyfeleiknek, mint hogy valaki amiatt ne vásároljon, mert elfogyott az iroda kínálata. Más szemszögből fordíthatnánk nagyobb figyelmet a nem sikeres adásvételek előrejelzésére is, hiszen inkább kevesebb bevételt tervezzen a cég, és utólag legyen mégis több, ám a könnyebbség és már meglévő kódolás miatt mi maradunk annál, hogy a sikeres adásvétel legyen a pozitív osztály.



23. ábra: Adásvétel előrejelzett valószínűségének eloszlása

A Probability-t használjuk fel a következő lépésben, ahol első körben 50%-os cut-value-t alkalmazva azokat a megfigyeléseket fogjuk az 1-es, azaz a sikeres adásvétel kategóriába sorolni, akikhez a Probability oszlop 50% vagy afölötti értéket számolt. Az előrejelzett valószínűségek alapján azonban látszódik ábránkon is, hogy nem az 50%, hanem inkább egy 20% körüli érték lesz jobb Recall(1) mutatót generáló cut-value, hiszen a mintaarány is ekörül mozog: a Maximum Likelihood, illetve az analógia elve mentén is a mintaaránnyal tudjuk legjobban közelíteni a sokasági arányt. Így második körben a mintaarány, 19,2%, lesz a cut-value (mely azt mutatja, hogy a mintában az ügyfelek 19,2%-a vásárolta meg a csomagot).

Valós (y)	Előrejelzett (y kalap)	
	0	1
0	3338	46
1	689	113

24. ábra: Klasszifikációs mátrix 50%-os Cut-value mellett

50% - os cut-value mellett tehát modellünk 159 sikeres adásvételt jelzett előre a 4186 megfigyelésből, lássuk, hogyan teljesít az alábbi mutatók tekintetében:

- Accuracy = 82,4%, ami a modell pontosságát fejezi ki a helyesen jelzett sikeres és sikertelen adásvételek összegének arányában az összes megfigyeléshez viszonyítva.
- Recall (1) = 14,1%, ami azt jelenti, hogy modellünk a valós sikeres adásvételek csupán 14,1%-át jelezte előre sikeresnek. Ez megegyezik a TPR mutatóval.
- Precision (1) = 71,1%, ami azt jelenti, hogy a modell által sikeres adásvételnek jelzettek közül 71,1% volt valóban sikeres.
- FPR (1) = 1,36%, ami azt jelenti, hogy a valójában nem sikeres adásvételek 1,36%-át azonosította modellünk sikeresnek.
- FNR (1) = 85,9%, ami azt jelenti, hogy a valójában sikeres adásvételek közül modellünk 85,9%-ot azonosított sikertelennek.
- TNR (1) = 98,64%, ami azt jelenti, hogy a valójában sikertelen adásvételek 98,64%-át jelezte előre sikertelennek a modell.

Ez a cut-value a sikeres kimenetel előrejelzése során tehát nem teljesít túl jól, inkább a sikertelent prediktálja helyesen és magas arányban. Ha csökkentjük a cut-value értékét, a Recall (1) mutató javítható (a Precision (1) romlásával párhuzamosan, de ez számunkra másodlagos), mivel ha 50%-nál kisebb valószínűség esetén is már sikeresnek fogja az adott megfigyelést prediktálni a modell, akkor összességében több sikeres előrejelzés várható, ami a Recall (1) számlálóját növeli (természetesen a több sikeres adásvétel előrejelzés nem mindegyike lesz helyesen sikeresnek jelezve, emiatt romlik a Precision, de mindenképpen a helyesen sikeresnek jel

zett metszet is nőni fog) a nevező változatlansága mellett, tehát a tört értéke is nagyobb lesz. D e nézzük meg, mennyivel javul a helyzet a mintaarány cut-value-nak választása esetén!

Valós (y)	Előrejelzett (y kalap)	
	0	1
0	2254	1130
1	291	511

25. ábra: Klasszifikációs mátrix mintaarány szerinti 19,2%-os cut-value mellett

Ezesetben modellünk valóban több, pontosabban 1641 sikeres adásvételt jelez előre, melyből 511-et helyesen, így a Recall (1) mutató feljavult 63,7%-ra, tehát modellünk a mintaaránybeli 19,2%-os cut-value mellett a valóban sikeres adásvételek 63,7%-át tudja előrejelezni, míg a Precision (1) leromlott 31,14%-ra, azaz a modell által sikeresnek jelzett adásvételek csupán 31,14%-a volt valójában sikeres.

Cut-value:	Accuracy:	Recall (1):	Precision (1):	TPR (1):	FPR (1):	FNR (1):	TNR (1):
50%	82,4%	14,1%	71,1%	14,1%	1,36%	85,9%	98,64%
19,2%	66,1%	63,7%	31,14%	63,7%	33,4%	36,3%	66,6%

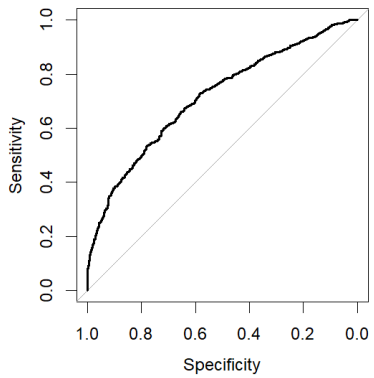
26. ábra: Mutatók összehasonlító táblázata

ROC – görbe

Modellünk ROC – görbéjéből visszaigazolódni látszik a gyenge-közepes magyarázóerő, hiszen bár felfelé húzódó jellege egyértelmű, azért messze áll a kívánt optimum ponttól (0% FPR és 100% TPR a bal felső sarokban helyezkedik el).

A görbe alatti terület nagysága is ezt támasztja alá: 0,7122 az eredményünk, melyet kontextusba helyez, hogy a maximálisan elérhető terület 1, a „vaktában találgató” modellé pedig ½, aminél modellünk 42%-kal jobb. Modellünk tehát valóban releváns, van magyarázóereje, de van még hova fejlődnie.

A ROC – görbe alapján is meg tudjuk adni az ideális cut-value értéket, amit jelen esetben mi a „closest.topleft” módszerrel határoztunk meg, melyre 18,77%-os cut-value értéket kaptunk, a ROC – görbe alapján is tehát a legmegfelelőbb a mintaarányhoz közeli értéket választanunk (ami 19,2% volt). Ezesetben a TPR 66%, a TNR pedig 65,2 %, azaz a FPR 34,8%. Ezek nagyon közeli értékek a mintaaránnyal kapott mutatókéhoz.



27. ábra: ROC-görbe

Előrejelzés perszónának

Modellünk gyakorlati felhasználásának jelentőségét egy konkrét példán keresztül is szemléltetnénk, létrehoztunk egy fiktív ügyfelet az alábbi adottságokkal:

- 25 éves kapcsolatban élő személy havi 20 000 dollár jövedelemmel, aki önként kereste fel az utazási irodát utazási csomag megvételét illetően egy 3 csillagos hotelbe, ahova 3-an utaznának, ezután meghallgatott egy 9 perces pitchet és további 3 alkalommal egyeztetett a feltételekről az iroda ügynökével.

Mit mond a modell? Modellünk perszónánk számára 18,58%-os adásvétel Odds-t jósol, ami mindhárom említett cut-value érték esetén még a 0-ás kategóriába sorolandó, tehát az ügyfél a modell szerint nem fogja megvenni az utazási csomagot. Kipróbáltuk, ha minden változót az Odds-növelő irányába mozdítunk el, akkor valóban megnő-e az adásvétel Odds. Így második, kicsit felturbózott perszónánk az alábbi:

- 23 éves egyedülálló személy havi 19 000 dollár jövedelemmel, akit irodánk keresett fel utazási csomag megvételét illetően egy 4 csillagos hotelbe, ahova 4-en utaznának, ezután ügyfelünk meghallgatott egy 12 perces pitchet és további 4 alkalommal egyeztetett a feltételekről ügynökünkkel.

Mit mond a modell? Modellünk feljavított perszónánk számára 55,1%-os adásvétel Odds-t jósol, ami mindhárom említett cut-value érték esetén az 1-es kategóriába sorolandó, tehát az ügyfél a modell szerint meg fogja venni a csomagot.

Nos, a modell előrejelzése megfelelő a paraméterek változtatása esetén, a két szélső esett között azonban számtalan olyan kevert alternatíva (ügyfél) létezik, akik esetén a prediktált adásvételi valószínűség 1-es vagy 0-as kategóriába való sorolása a cut-value értékétől függ (azaz, ha valaki 18,77% és 50% közé esik, ott eltérő a 3 cut-value-tól függően a besorolása).

Összegzés

A dokumentumban tehát ismertettük a használt adatbázisunkat, elvégeztük az adattisztítási lépéseket és egy alap leíró statisztika bemutatásával és a változók vizualizálásával előkésztettük a munkánkat a későbbi feladatrészekre. Ezek után egy regressziós modellt építettünk, amelyet modellszelekcióval tökéletesítettünk, hogy a lehető legpontosabb becsléseket sikerüljön eredményül kapnunk. Az eredményeket többféleképpen értékeltük, értelmeztük és következtetéseket vontunk le belőlük. Utolsó lépések között elvégeztük a klasszifikációs mátrix, az abból számolandó mutatók értelmezését és a ROC-görbe vizualizációját. Végül a gyakorlati felhasználás példájaként elkészítettünk egy előrejelzést két általunk létrehozott perszónának, hogy a modell alapján megvalósulna-e a sikeres adásvétel vagy sem.

Források:

Kaggle:

Utazás előrejelzése adatbázis:

Letöltés helye: https://www.kaggle.com/datasets/susant4learning/holiday-package-purchase-prediction/data [Utolsó letöltés ideje: 2023. 11. 14.]