1. Bevezetés

Elemzésünk alapját a Kaggle.com oldalról letöltött, 5000 legpopulárisabb filmet tartalmazó adatbázis képezte („The Movies TMBD"). A táblázatot megtisztítottuk az üres cellákat tartalmazó soroktól, majd a további hibás sorokat is kiszűrtük (pl.: revenue=0), így eljutva egy 2149-es minta elemszámhoz. Dolgozatunkban bizonyos részeinél az egész adatbázist felhasználtuk (pl.: logisztikus regresszió), míg más helyen csak egy kisebb részével dolgoztunk (pl.: Walt Disney filmek) a feladat igényeihez igazodva. Mivel mindketten nagy filmrajongók vagyunk, kíváncsian vártuk, hogy az általunk kigondolt hipotézisek vajon statisztikai módszerekkel is igazolhatók-e (pl.: animált filmek tényleg olcsóbbak-e, rövidebbek, stb.).
Dolgozatunkban felhasznált változók:

- title: filmcím (eredeti nyelven), kategórikus
- budget: költségvetés (USD, bruttó, 2010-es árfolyamon inflációval korrigálva), folytonos
- revenue: bevétel (USD, bruttó, 2010-es árfolyamon inflációval korrigálva), folytonos
- popularity: IMDB által számolt népszerűségi érték (felhasználói aktivitás alapján), folytonos
- runtime: játékidő (perc), folytonos
- vote_average: átlagos IMDB pontszám (1-10 skálán), folytonos
- vote_count: IMDB szavazatok (db), folytonos
- original_language_en: eredeti nyelv angol, kategórikus
- num_prod_countries: a filmben szereplő országok (db), kategórikus
- num_spoken_languages: a filmben felhangzó nyelvek (db), kategórikus
- num_bigsix: a 'Big Six'-be tartozó producer vállalatok száma, kategórikus
- animation: animációs (vagy élő szereplős), kategórikus
- action: akció, kategórikus
- comedy: vígjáték, kategórikus
- drama: dráma, kategórikus
- thriller: thriller, kategórikus
- age: a film kora években (a referencidátumhoz viszonyítva - 2017/09/28), folytonos

elemzés neve	az elemzésben szereplő változók neve	van-e eredményváltozó az elemzésben (ha igen, akkor melyik változó az, és milyen mérési szintű az eredményváltozó)
klaszterelemzés	7db folytonos	
kereszttáblás elemzés		la_an
faktorelemzés	7db folytonos	
logisztikus regresszió	először:7db folytonos és 13 db kategórikus, majd szűrés után: 4db folytonos, 1 db kategórikus	animation(animation,live)

2. Klaszterelemzés

A klaszterelemzéshez az eredeti adatbázisunk sajnos túl nagy lett volna, kevésbé tudtuk volna a nagy mintaszám miatt a dendogrammokat és az egyéb kapott eredményeket értelmezni, ezért leszűrtük az adatokat, hogy csak azokat a filmeket tartalmazza, amelyeknek a Walt Disney egyedüli vagy társ producer volt. Így egy bizonyos szempontból sokkal homogénebb, 77 elemű mintát tudtunk elemezni adatbázisunk sztenderdizált folytonos változói alapján (revenue, budget, popularity, runtime, vote_average, vote_count, age, melyek értelmezése az első fejezetben megtalálható).

2.1. Klaszterkönyök

A klaszterelemzést a szűkített mintán az optimális klaszterszám megállapításával kezdtük, amihez először is a K-középpontú („K-means”) módszerrel a klaszterkönyököt vizualizáltuk. A hüvelykujj-szabály alapján maximum 6 klaszter lehet indokolt (~ $[77/2]^{1/2}$), ezért 2-től 6-ig először lefuttattuk és kimentettük a klaszterbesorolásokat a K-means módszerrel, amelyeket azután egyesével faktorként használva egyirányú ANOVA táblákat készítettünk (Melléklet 2.1.1.). Ezekből a külső és teljes variancia-összegek arányát vizsgálva kirajzolhattuk a klaszterkönyök megállapítására szolgáló ábrát (Melléklet 2.1.2.). Ezen már szabad szemmel is kivehető a klaszterkönyök a 4-es klaszterszámnál, itt látható szignifikáns törés a görbe meredekségében, azaz a klaszterszámok növelésének ezután már nincs akkora hozzáadott értéke.

2.2. K-középpontú klaszterelemzés

Mindezek után újra lefuttattuk a K=4 optimális klaszterszámmal a K-középpontú metódust a fent felsorolt sztenderdizált folytonos változóinkra. A végleges klaszterközéppontoktól vett távolságokat és a klasztertagságok számát tartalmazó táblázatokat (Melléklet 2.2.1.) értelmezve a következők mondhatók el klaszterjeinkről (átlag alatt a központokat, 'mean'-eket értjük):

- 1-es klaszter: Ez egy egyelemű klaszter lett, egyetlen eleme pedig a Fantasia című film. Elsőre meglepő lehet egy egyelemű klaszter léte, de jobban belegondolva nem olyan nagy csoda ez, hiszen a film 1940-es (míg a maradék legtöbb film korunkbeli), ami nemhogy önmagában egy outlier „age" változót eredményez, de a korbeli különbség a többi változóra is természetes módon rányomja bélyegét (átlagnál kisebb költség, kisebb bevétel, kisebb popularitás, stb.).
- 2-es klaszter: Ide került a filmek többsége, melyekről elmondható, hogy átlagos vagy inkább az átlagnál picit gyengébb filmek. Az, hogy valamennyi pozitívan értelmezhető mutatónál az átlagosnál picit gyengébb eredményt értek el, talán betudható annak is, hogy az átlagnál régebbi filmekről beszélünk.
- 3-as klaszter: Szintén egy relatív népes klaszterről beszélhetünk, ahol a táblázatot vizsgálva azt láthatjuk, hogy az átlagnál jobb minőségűnek (vote_average és vote_count alapján) tartott filmek kerültek ide, amelyek ráadásul az átlagosnál kicsivel több büdzséből az átlagosnál viszonylag nagyobb bevételt tudtak eredményezni.
- 4-es klaszter: Ebbe a pár elemű klaszterbe kerültek a tényleg viszonylag drága előállítási költségű, felkapott (popularity alapján), újabb és hosszabb filmek, amelyek az átlagnál több bevételt is produkáltak.

Az elemzésnél lekért ANOVA tábla megmutatta (Mellélet 2.1.1. 3. táblázata), hogy mindegyik felhasznált változó szignifikáns volt a klaszterek kialakításában. Fontos kiemelni még továbbá, hogy ezek csupán a Walt Disney filmek a már eleve korlátozott elemszámú mintánkból. Ennek ellenére a 4 klaszter interpretációja rendkívül érdekes és gondolatébresztő, ezért úgy döntöttük ezeket használjuk majd fel a kereszttáblás elemzésben.

2.3. Hierarchikus klaszterelemzés

A hierarchikus klaszterelemzést két módszerrel, a Ward („Ward Linkage") és átlagos („Average linkage between groups") módszerekkel elemeztük. Megfigyelésünk központjában a két

módszerrel kapott dendogrammok álltak, mely ábra arra szolgál, hogy vizualizálja, hogy a program egy bizonyos meghatározott elv alapján milyen sorrendben osztja be, rendezi az elemeket a különböző klaszterekbe. A dendogrammokat nagyságuk miatt a mellékletben csatoltuk (Melléklet 2.3.1. a és b), az ábrákat megfigyelve tökéletesen kivehető a két eljárás közti különbség mivolta.

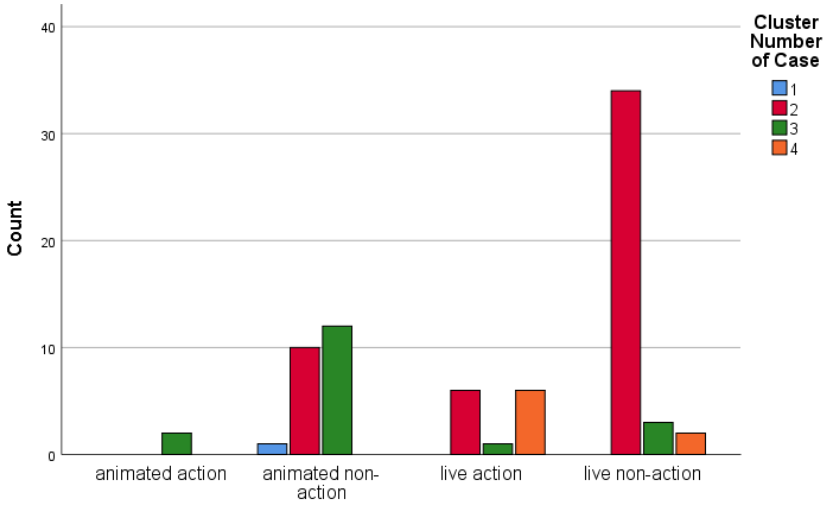
A Ward Linkage módszer az egyik legnépszerűbb és igen nagy előszeretettel alkalmazott módszer, melynél a klaszterek összevonása egy előre meghatározott funkció alapján, a mi esetünkben a varianciák növekedésének minimalizálásával végzi el a program az agglomerációt. Az Average Linkage módszer két klaszter távolságát úgy határozza meg, hogy párosával vett távolságok átlagát elemzi, próbálva ezt minimalizálni. A legközelebbi és legtávolabbi szomszéd eljárásokkal ellentétben ez az összes páros távolságot figyelembe veszi, ezek átlagát veszi.

Ezekből a tulajdonságaikból fakad, hogy melyik klaszterbe és hányadik lépésben emelnek be egyes elemeket az eljárások, de ennek ellenére az elemek majd ¾-ét ugyanabba a klaszterbe feltételezték (és megnyugtatóan a már említett Fantasia filmet mindkét módszer egy egyelemű külön klaszternek vélte). A Ward linkage dendodgramm egy körülbelül szimmetrikus faábrát mutat, ahol már a vártakkal megfelelően azt láthatjuk, hogy a program a hamar kialakított, több kisebb csoportot igyekszik folyamatosan egyre bővíteni és ezeket összevonva nagyobb klasztereket kreálni. Ezzel szemben az Average linkage módszernél egy elnyúló ábrát figyelhetünk meg, ami azt jelenti, hogy sok utolsó körös kis elemszámú, addig hanyagolt klaszterek bevonása is történt főképp a metódus vége felé. Ha tehát vágással kéne megállapítanunk csupán a dendogrammokat vizsgálva a 4 klasztert, akkor Ward módszerrel kisebb távolságra lévő klasztereket kapnánk, mint az Average-nél.

3. Kereszttáblás elemzés

A K-középpontú klaszterelemzésnél a középponti távolságokat értelmezve érdekesnek ígérkező csoportokat kaptunk, így ezekre szerettük volna elvégezni a kereszttáblás elemzést, hátha valamilyen magyarázatot kapunk a csoportok tényleges mivoltára. Feltételeztük, hogy valamilyen szinten, az közrejátszhat a csoportosításban, hogy egy film milyen besorolású, milyen korosztály a célközönség, ugyanis ez mind közrejátszhat a bemeneti változóink milyenségében (film hossza, bevétele, költsége, stb…). Arra gondoltunk, hogy az körülbelül definiálhatja a célközönséget, hogy egy film akció-e vagy nem, valamint hogy élőszereplős-e vagy animációs. Erre volt is

adatunk, így ezek kombinációit (értelemszerűen négy van) használtuk fel (az la_an új kategorikus változót kreálva) a K-középpontú klaszterekkel összevetve a kereszttáblás elemzésben, melyre a következő ábrát kaptuk eredményül (a táblázat a mellékletben megtekinthető):



Természetesen elvégeztük a Chi-négyzet tesztet is elvégeztük, amely eredménye alább megtekinthető.

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	47,251[a]	9	,000
Likelihood Ratio	41,589	9	,000
N of Valid Cases	77		

a. 11 cells (68,8%) have expected count less than 5. The minimum expected count is ,03.

A szinte 0 p érték alapján, a H0 hipotézist, miszerint a két kategorikus változónk tökéletesen függetlenek a populációnkban, elvethetjük. Azonban az SPSS többek között még egy fontos feltevést tesztelt futásnál, miszerint a táblánk nem több mint 20%-ának a várt gyakoriságának kevesebbnek kell lennie, mint 5 (nagyobb táblákra vonatkozó hüvelykujj szabály alapján). Nálunk ez 68,8% lett, amely alapján nem támaszkodhatunk biztonsággal e teszt eredményére. Ez egyébként magyarázható azzal, hogy pl. a klasztereink között is fellelhető egy egyelemű (outlier értékű) klaszter, mely belegondolva nem igazán összeegyeztethető az elgondolt kereszttáblás összehasonlítási kategóriáinkkal.

Ennek ellenére ezután megvizsgáltunk még egyéb mutatókat is az alábbi táblázat alapján:

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,783	,000
	Cramer's V	,452	,000
	Contingency Coefficient	,617	,000
N of Valid Cases		77	

Ez alapján mindhárom vizsgált index szignifikáns jelentőségű, azonban mindhárom a Chi-négyzet teszten alapul, így nem lehetünk biztosak az általuk közölt eredmény alkalmazhatóságában. A kontingencia koefficiens, a Phi és a Cramer's V mutató is 0-1 intervallumon értelmezhető (0 a tökéletes függetlenséget, míg 1 a függvényszerű kapcsolatot jelenti), és mindhárom alapján egy közepes asszociációt feltételezhetünk a két kategorikus változónk között.

A fent tárgyalt alkalmazhatósági korlátok tudatában az ábrát és az eredményeket vizsgálva elmondhatjuk, hogy a 2-es klaszterbe sorolt átlagos és átlagnál picit gyengébb filmek (valamennyi változó alapján) legtöbbje nem akció, élőszereplős film. Tehát ez alapján a 2-es klaszter filmjeinek legtöbbje bizonyára valamilyen (Disney) középkategóriás dráma vagy vígjáték, azaz egy kimondott „szombat estés családi TV film", amelyekről tapasztalatból elmondható, hogy kb. közepes a büdzséjük, bevételük, hosszuk, stb.. A 3-as klaszter elemei közül a legtöbb a nem akció animációs film kategóriába került. Ez a kategória a Disney vezérhajója, az animációs gyerekfilmek piacát egyértelműen a Disney uralja már évtizedek óta, így összeegyeztethető, hogy e filmek azok, amikre a Disney sokat erőforrást allokál, de a bevételek is magasak, valamint a széles közönség tetszését is elnyerik. Érdekes még, hogy az összes animált akciófilm a 3-as klaszter elemei közül kerül ki. A maradék kategóriákkal és klaszterekkel már nem volt ilyen szerencsénk, nem igazán lehet releváns megállapításokat levonni a kereszttáblás elemzésünk alapján.

4. Faktoranalízis (PCA és PAF)

A faktoranalízist PCA (Principal Component Analysis=főkomponens-elemzés) és PAF (Principal Axis Factoring) Dimension Reduction módszerekkel végeztem el. Mivel korrelációs mátrixot használtunk mindkét módszernél, mindegy hogy sztenderdizált vagy nem sztenderdizált

adatokkal dolgozunk. Az elemzésbe az összes fent említett folytonos változónkat bevontuk (revenue, budget, popularity, runtime, vote_average, vote_count, age).

4.1. Elvégezhetőség vizsgálata

A vizsgálati praktikák a két módszer esetében megegyeznek. Először a változók közti korrelációs mátrixot vettük szemügyre. A páronként korrelációkról elmondható, hogy körülbelül mindenhol legalább gyengén közepes erősségű, néhol erős közepes. A mátrix determinánsa 0,048, ami 0-hoz közeli érték révén azt indukálja, hogy az adatok megfelelnek a főkomponens-elemzésre (Melléklet 4.1.1.). Ezt a tényt, alátámasztja a KMO teszt 0,734-es egyhez közeli értéke, valamint a függetlenséget H0 hipotézisként feltételező Bartlett teszt eredménye is (a p érték szinte 0, így H0-t elvethetjük, az adatok nem függetlenek egymástól) (Melléklet 4.1.2.). Az elvégezhetőséget és a változók közti kapcsolat erősségét magyarázza még az is, hogy az anti-imázs mátrix főátlójában csak 0,5 feletti (1-hez közeli) értékek szerepelnek (Melléklet 4.1.3.).

4.2. A módszerek összehasonlítása

A PCA módszer lineárisan állítja elő a bemeneti változókból a komponenseket, úgy hogy az egyedi varianciát maximalizálja, azaz a minél nagyobb különbözőségre törekedve, így magyarázva a komponensekkel a változók varianciáját.

A PAF módszerrel a változók közti korrelációt a komponensekkel magyarázzuk, azt feltételezve, hogy vannak ugyan látens változók, de azokat közvetlenül mérni nem tudjuk. Ehhez csökkentett korrelációs mátrixokat használ az elemzés, melyben a diagonálisok az alapvető 1-esek helyett kommunalitások (azaz varianciát magyarázó mérőszámok). Mivel a kommunalitások azok, amiket keresnénk a módszerrel, ezért iterálva kezdjük el a csökkentett korrelációs mátrixnak az előállítását az 1-es diagonálisú korrelációs mátrixból indulva. Lényegében tehát a diagonálist lecserélve egészen addig iterálunk folyamatosan közelítésekkel, míg a kommunalitások már nem változnak egy-egy futás között szignifikánsan.

4.3. Az eredmények összehasonlítása

A PCA esetén végül minden változó esetében kb. legalább ugyanakkora, de legtöbbször nagyobb kommunalitásokat vélhetünk felfedezni. PCA alapján egyik változó kizárása sem indokolt, PAF

esetén az age és a runtime változókon már erősen el lehetne gondolkozni, de az összehasonlíthatóság jegyében most bent tartottuk e változókat (Melléklet 4.3.1.).

A rotált komponens mátrix és a rotált faktor mátrix (Melléklet 4.3.2. és alább is megtekinthető) is hasonló irányú (és legtöbbször hasonló nagyságrendű) értékeket mutat (bár a faktormátrix a fentebb említett változók benntartása miatt kevésbé meggyőző önmagában), két látenst azonosított mindkettő módszer. A teljes varianciát magyarázó táblák (Melléklet 4.3.3.) alapján látható, hogy a PCA két látense 66% varianciát magyaráz összesen (ebből az első komponens 43%), míg a PAF-nál a benntartott változók miatt a látensek által magyarázott kumulatív rotált variancia értéke 54% (ebből az első faktor 37%). Visszatérve a rotált mátrixok értékeinek vizsgálatára elmondható, hogy az első látens húzó változói a pénz típusúak (budget és revenue) és a népszerűséget kifejezőek (vote_count és popularity, azaz „hype"). A második látensbe, ha egy közös elnevezést kéne adnom neki, a film minőségi ismérveit kifejező változók kerültek (age, runtime, vote_average, PAF esetén még a vote_count és popularity változóknak is szignifkánsnak mondható jelenléte van itt is).

Rotated Component Matrix[a]

	Component 1	2
budget	,797	-,129
revenue	,888	,105
age	-,408	,613
popularity	,792	,162
runtime	,211	,655
vote_average	,187	,802
vote_count	,879	,243

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 3 iterations.

Rotated Factor Matrix[a]

	Factor 1	2
budget	,738	-,073
revenue	,875	,150
age	-,313	,287
popularity	,690	,239
runtime	,161	,381
vote_average	,076	,875
vote_count	,847	,347

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 3 iterations.

5. Logisztikus regresszió

Dolgozatunk ezen részében a teljes, 2149 elemű adatbázissal dolgoztunk. Célváltozónak (függő változó) az „Animation" bináris változót választottuk, melynek a (0,1) paramétereihez az „Animation" és a „Live" értékek vannak hozzárendelve, ami azt jelenti, hogy az adott film animációs technikával készült, vagy sem. Legelőször, a korábban bemutatott összes lineáris

változónkat standardizáltuk, hogy javítsunk a sokaság összehasonlíthatóságán (5.1. Melléklet). Ezután az összes változónk (15 db) bevonásával lefuttattuk a logisztikus regressziót a Forward Wald módszer szerint. Következő lépésnek egyesével megvizsgáltuk a változóink eredményül kapott szignifikancia szintjét, és amelyek 0,01-nél nagyobbak voltak, azokat kivettük a modellből. (5.2. Melléklet). Így csak azok a változók maradtak a modellben, amelyek 1%-os szignifikancia szinten tényleges magyarázó erővel bírnak. A multikollinearitás kiszűrése céljából pedig a kapott 5 változót a Pearson féle korrelációval is elemeztük. Hüvelykujj szabály szerint a 60%-os korrelációs szintet tekintettük elvetendőnek, azonban az 5 változónknál egyik sem lépte át ezt a szintet, így az összes benne maradhatott a modellünkben (5.3. Melléklet).

- budget: (költségvetés (USD, bruttó, 2010-es árfolyamon inflációval korrigálva)
- revenue: bevétel (USD, bruttó, 2010-es árfolyamon inflációval korrigálva)
- runtime: játékidő (perc)
- vote_average: átlagos IMDB pontszám (1-10 skálán)
- drama: dráma típusú volt-e a film, vagy sem (Drama, not)

A magyarázó változók együttes inszignifikanciáját elvethetjük az Omnibus teszt alapján (5.4. Melléklet). A Wald teszt alapján a változók egyesével, külön-külön is szignifikánsnak mondhatóak. (5.5 Melléklet). A "Cox & Snell R2" értéke 0,252, míg a "Nagelkerke R2" értéke pedig 0,667 (Melléklet 5.6), vagyis csak a konstanst tartalmazó null-modell log likelihood értéke ennyivel csökkent a magyarázó változók bevonása miatt. Mivel 0,35 felett a pseudo R^2-ek értéke megfelelőnek tekinthető, ezért a mi esetünkben lévő 0,667 már egészen kiválónak mondható. A Hosmer and Lemeshow teszt a modell illeszkedését vizsgálja. A mi esetünkben a p érték 5% alatt van (5.7. & 5.8 Melléklet), ezért el kell utasítanunk a H0 hipotézist, miszerint a becsült és tényleges valószínűségek között nem szignifikáns az eltérés, vagyis nem megfelelő a modell illeszkedése.

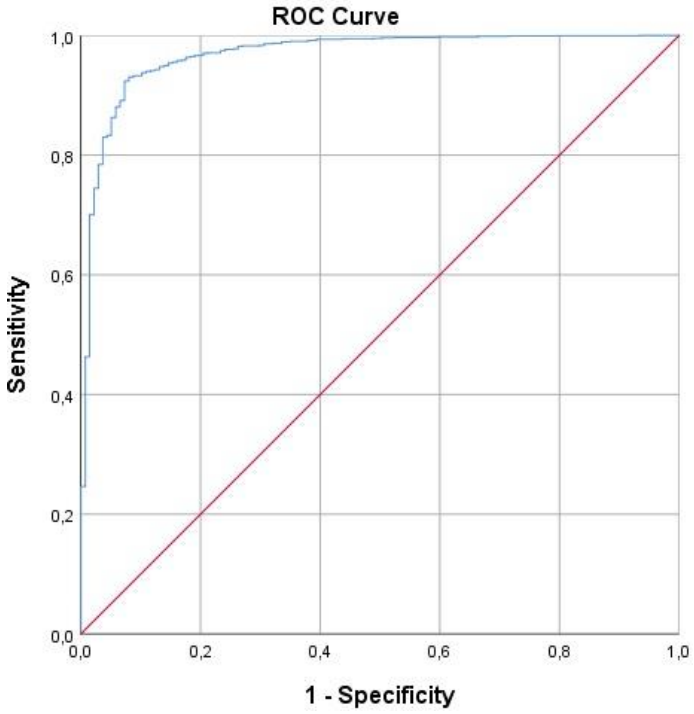
A logisztikus regressziónk egyenlete (Melléklet 5.5) a következő lett:

$$odds=(p/1-p)=8{,}181*\text{drama}+0{,}233*\text{budget}$$
$$+0{,}665*\text{revenue}+85{,}800*\text{runtime}+0{,}363*\text{vote\_average}$$
$$p=1/(1+e-(5{,}923+2{,}102*\text{drama}-1{,}455*\text{budget}-0{,}408*\text{revenue}+4{,}452*\text{runtime}-1{,}014*\text{vote\_average}))$$

A modell alapján a következő értelmezés igaz például a budget változóra: Amennyiben egy egységgel (1USD) növekedik a budget ceteris paribus, akkor 0,233-szeresére növeli az oddsot ((p/(1-p) hányadost), vagyis azt, hogy az adott film animált, vagy élő szereplős. Jelen esetben a

0=animált, és az 1=élőszereplős, vagyis a film költségvetésének növekedése növeli annak az esélyét, hogy a film élőszereplős lesz, ami egybe vág az előfeltevéseinkkel. Ugyanis egy animált film elkészítése véleményünk szerint a legtöbb esetben magasabb költségekkel jár (drága színészek, egzotikus forgatási helyszínek, stb.). A többi változót is megvizsgálva láthatjuk, hogy szintén teljesülnek az elővárakozási feltevéseink (pld.: animált filmek ált. nem drámák, kevesebb bevételt hozhatnak, rövidebbek, stb.).

A logisztikus görbénk megfelelőségének vizsgálatára a ROC görbe alatti terület mérése (ROC AUC) is egy jó módszer. Az a modell számít jónak, amelynél a ROC görbe minél jobban rásimul a négyzet bal felső sarkára, ezáltal növelve a függvény alatti terület értékét. Modellünkben 0,968 lett ez az érték, ami 0,8-as hüvelykujj szabály fölött található, így egészen kiválónak mondható (5.9. Melléklet).



6. Diszkriminancia elemzés

Diszkriminancia elemzés célja, a csoportok szétválasztása a kanonikus térben. Azért, hogy a kapott eredmények jól összehasonlíthatóak legyenek az előző feladatban kapott értékekkel, itt is az „Animation” Dummy változó értékeit vizsgáltuk. Mivel az előző feladatban is jó magyarázó erővel bírt a választott 5db magyarázó változónk az „Animation” kimenetelére, ezért az lett az

előfeltevésünk, hogy itt is jól el fog különülni egymástól a két csoport, az animált és az élőszereplős filmek.

6.1 Előfeltevések

A diszkriminancia alkalmazásának több előfeltétele van. Az egyik, hogy csoportonként a változók átlagai legyenek eltérőek. Ezt a feltételt a Wilks' Lambda teszttel vizsgálhatjuk, melynek nullhipotézise, hogy a csoportátlagok nem különbözőek. Az összes változónkra egy 0 közeli p értéket kaptunk, így H0-t elvetettük, és a H1 hipotézist fogadtuk el (Melléklet 6.1.1.). Másik feltétele az alkalmazásának, hogy a változók eloszlása legyen többdimenziós normál eloszlás. A normalitás teszt elvégzése után mind az 5 változónkra 0 közeli p értéket kaptunk, melynek következtében el kellett vetünk a nullhipotézist, miszerint az adott változó az adott pontban normális eloszlást követ. (Melléklet 6.1.2.). Ezt követően mind az 5 változót át transzformáltuk logaritmus alapú változóra, majd újra lefutattuk a normalitás tesztet, azonban így is el kellett vetnünk a nullhipotézist (Melléklet 6.1.3.), ezért az eredeti változóinkkal folytattuk a számolást. Végül pedig a Box's M tesztel azt vizsgáltuk, hogy a csoport kovariancia mátrixok megegyeznek-e. A teszt nullhipotézise alapján a variancia-kovariancia mátrixok nem különbözőek, melyet a kapott értékeink alapján el kellett vetnünk, vagyis a variancia-kovariancia mátrixok különbözőek.(Melléklet 6.1.4.).

6.2 Számítások

Miután lefuttattuk a diszkriminancia elemzést láthattuk, hogy nincsenek hiányzó elemek, mind a 2149 film belekerült a modellbe.(Melléklet 6.2.1) Ezt követően a Group Statistics adatait átmásoltuk egy excel táblába, ahol a mintaelem szórásokat elosztottuk a mintaelem átlagokkal, így megkapva a relatív szórást. Hüvelykujj szabály szerint a relatív szórásnak egy 2-nél kisebb értéket kell lennie, különben az „outlier"-ek ronthatják a modellt. Példánkban csak az animált megjelenítésű drámáknál volt egyetlen kiugró érték. (Melléklet 6.2.2.).

A kanonikus korreláció azt mutatja meg, hogy mennyire szoros az asszociáció a kapott diszkriminancia értékek (mint függő változók) és a csoportok között. Azt méri, hogy a diszkrimináló értékek változékonyságát milyen arányban magyarázza a csoportbesorolás. A kanonikus korreláció mindig egy 0 és 1 közé eső szám, amelynél a minél nagyobb érték a jobb, mivel ilyenkor erősebben a korreláció. A mi példánkban ez 0,49 lett, ami jónak mondható. (Melléklet 6.2.3.).

A módszer alkalmazhatóságának feltételeinél teszteltük az egyes változók Wilks' Lambda értékét, azonban meg kell vizsgálnunk a függvények (és nem az eredeti változók hatásait is). A Wilks' Lambda alapvetően a csoportokon belüli átlagos négyzetes eltérés és teljes átlagos eltérés arány, amely a diszkrimináló függvény minőségét adja meg. Értéke itt is mindig egy 0 és 1 közé eső szám, ahol a 0-hoz közelítő értékek azt jelentik, hogy a csoportokon belüli variabilitás kicsi, vagyis a függvényünk jól diszkriminál a csoportok között. Ezzel ellentétesen, ha 1-hez közelít az értéke, akkor a függvény kevésbé tudja a csoportokat elkülöníteni egymástól. Számításunk során egy magas, 0,760-os értéket ad eredményül, amelyből látszik, hogy nem tudja olyan jól elkülöníteni a csoportokat, amely logikusan következik diszkriminancia elemzés előfeltevéseinek részleges hiányából. (Melléklet 6.2.4.). A diszkrimináló függvény egyenletét megkaphatjuk a koefficiensek táblázatból (Melléklet 6.2.5.). Példánkban a függvény egyenlete az alábbi:

*0,242*drama+0,744*budget+0,186**revenue-*0,943*runtime+0,490*vote_average+0,109=0*

6.3 Összehasonlítás a logisztikus regresszióval

Látható, hogy a diszkriminancia előfeltevéseiből kettő nem teljesült, ezért nem igazán lehet teljes bizonyossággal értelmezni a fenti tesztek eredményeit. A logisztikus regresszió a leginkább hasonló eljárás a diszkriminancia analízishez, mivel ez a módszer is egy kategórikus függő változó magyarázatára szolgál. A logisztikus regresszió többek között abban az esetben is választandó eljárás, ha a fent említett alapfeltevések nem teljesülnek, ugyanis ezekre a feltételekre a logisztikus regresszió nem érzékeny, ezért jelen esetben az ott kapott eredmények pontosabb értékeket adnak.

Összegzés

Ahogy a bevezetőben is említettük, sikerült egy számunkra (és reméljük másoknak is) nagyon élvezetes és sokszínű adatbázist választani. A végrehajtandó elemzések során kapott eredményeket érdekesnek és izgalmasnak találtuk. Amennyiben lesz rá lehetőségünk, akár más tantárgyak keretében is szívesen foglalkoznánk a témával a jövőben, ugyanis sok érdekes előfeltevésünkre kaptunk olykor elgondolkodtató válaszokat.