



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSYLVANIAN STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Hitelminősítő algoritmusok összehasonlítása: mesterséges intelligencia alapú algoritmusok vs. hagyományos módszerek

Szilágyi-Nagy Zétény, III. év, alapképzés,
Babeș-Bolyai Tudományegyetem, Közgazdaság- és Gazdálkodástudományi Kar

dr. Dézsi-Benyovszki Annamária, egyetemi docens,
Babeș-Bolyai Tudományegyetem, Közgazdaság- és Gazdálkodástudományi Kar



Összefoglaló

A tanulmány célja összehasonlítani a logisztikus regresszió, döntési fa és a véletlen erdő teljesítményét egy valós (peer-to-peer, P2P)¹ hitelezési adatbázis alapján. A peer-to-peer (vagy személytől személyhez szóló, P2P) hitelezési piac megkönnyíti a hitelfelvevők és a hitelezők közötti pénzügyi tranzakciókat. A hitelkockázat kezelés hagyományosan a hitelpontozáson alapszik, amely segítségével becsüljük a hitelek nemteljesítésének valószínűségét. A megfelelő hitelminősítési technika nagyon fontos úgy a pénzügyi intézmények, mint a P2P hitelezési platformok hosszú távú sikere szempontjából.

A hitelminősítő algoritmusok összehasonlítása a Lending Clubról származó adatok alapján történik. A kapott eredmények arra a következtetésre engednek jutni, hogy a logisztikus regresszió a legjobb hitelminősítő algoritmus a Lending Club adatbázisa esetében.

¹ Peer-to-peer hitelezés egy platformalapú vállalkozás, ami egy központi piacot hoz létre, ahol a hitelfelvevők és a hitelezők találkozhatnak. A platformok a hagyományos közvetítői szegmens kihagyásával a bankok ajánlatainál kedvezőbb hitelfelvételi lehetőséget ígérnek a finanszírozást igénylő kis ügyfeleknek, illetve a befektetni vágyók számára – bár nagyobb kockázat mellett – a bankbetétnél magasabb kamatot kínálnak (Dömötör és Ölvedi, 2023).



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSYLVANIAN STUDENTS' SCIENTIFIC CONFERENCE

ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Tartalomjegyzék

Bevezető	1
Szakirodalmi áttekintés	3
Adatok és módszertan.....	9
Az adatok.....	9
Módszertan	16
Eredmények.....	19
3.1. Változók közötti kapcsolatok elemzése	19
3.2. Változók kategorizálása és a WoE érték kiszámítása	25
3.3 Tanuló és teszt adatbázis	25
3.3.1. Logisztikus regresszió becslése	26
3.3.2. Döntési fa.....	31
3.3.3. Véletlen erdő.....	37
Irodalomjegyzék.....	42



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Ábrák és táblázatok jegyzéke

Ábrák:

1. Ábra: A Lending Club által folyósított hitelek száma, 2016-2022.....	9
2. Ábra: Az ügyfelek régiség szerinti eloszlása	14
3. Ábra: Az ügyfelek jövedelem szerinti eloszlása	14
4. Ábra: A hitelösszeg átlaga államok szerint.....	15
5. Ábra: Az ügyfelek minősítés szerinti eloszlása.....	16
6. Ábra: Az ügyfelek nemteljesítés és futamidő szerinti eloszlása	19
7. Ábra: Az ügyfelek nemteljesítés és minősítés szerinti eloszlása.....	20
8. Ábra: Az ügyfelek nemteljesítés és az al-minősítés szerinti eloszlása.....	21
9. Ábra: Az ügyfelek nemteljesítés és a jelenlegi munkaviszony hossza szerinti eloszlása.....	21
10. Ábra: Az ügyfelek nemteljesítés és a lakástulajdon szerinti eloszlása	22
11. Ábra: Az ügyfelek nemteljesítés és az ellenőrzési állapot szerinti eloszlása	22
12. Ábra: Az ügyfelek nemteljesítés és a hitel célja szerinti eloszlása.....	23
13. Ábra: Az nemteljesítés és a lakhely szerinti eloszlása	23
14. Ábra: ROC görbe: logisztikus regresszió	28
15. Ábra: ROC görbe: logisztikus regresszió WoE transzformációval.....	30
16. Ábra: Metszett alulmintázott döntési fa	31
17. Ábra: ROC görbe: döntési fa alulmintázással	33
18. Ábra: ROC görbe: döntési fa veszteségmátrix bevonásával.....	35
19. Ábra: Döntési fa előzetes valószínűségek megváltoztatásával	36
20. Ábra: ROC görbe: döntési fa előzetes valószínűségek megváltoztatásával.....	37
21. Ábra: ROC görbe: véletlen erdő	38
22. Ábra: ROC görbék: összehasonlítás	39

Táblák:

1. Táblázat: Hitelek száma az átalakítások előtt és után	10
2. Táblázat: Lending Club-ról származó változók	11
3. Táblázat: A numerikus változók leíró statisztikái.....	13
4. Táblázat: Mennyiségi változók és nemteljesítés közötti kapcsolat.....	24
5. Táblázat: A nemteljesítési valószínűséget befolyásoló tényezők a logisztikus regresszió alapján.....	26
6. Táblázat: Klasszifikációs mátrix: logisztikus regresszió.....	28
7. Táblázat: A nemteljesítési valószínűséget befolyásoló tényezők WOE transzformációval ..	29
8. Táblázat: Klasszifikációs mátrix: logisztikus regresszió WOE transzformációval.....	30
9. Táblázat: Klasszifikációs mátrix: alulmintázott döntési fa	32
10. Táblázat: Klasszifikációs mátrix: döntési fa a veszteségmátrix bevonásával.....	34
11. Táblázat: Klasszifikációs mátrix: döntési fa előzetes valószínűségek megváltoztatásával	36
12. Táblázat: Klasszifikációs mátrix: véletlen erdő	38
13. Táblázat: Összefoglaló táblázat mutatókkal	39



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Bevezető

Teply és Polena (2020) szerint a FinTech pénzügyi innovációjaként a személyközi (peer-to-peer, P2P) hitelezés egy új, online alapú pénzügyi közvetítés, amely a hitelfelvételre hajlandó embereket (hitelfelvevőket) köti össze a pénzüket kölcsönadni hajlandó emberekkel (hitelezőkkel). A hitelfelvevők és a hitelezők online P2P hitelezési platformokon keresztül kerülnek kapcsolatba egymással. A P2P hitelezési platformok a hagyományos bankoknál alacsonyabb közvetítési költséggel tudnak hiteleket nyújtani, mivel a működési költségeik alacsonyabbak, mint a fizikai bankoké az online jelenlétük miatt. Ez támogatja a versenyképesebb feltételek kínálását mind a hitelfelvevők, mind a hitelezők számára. A hitelfelvevők átlagosan alacsonyabb kamatlábakat fizetnek a P2P-hitelezési platformokon, mint a bankoknál (Namvar, 2013). Hasonlóképpen, a jól diverzifikált hitelportfólióval rendelkező hitelezők magasabb hozamot érnek el, mint a hagyományos megtakarítási számlákon (Serrano-Cinca et al., 2015). Ezek a mögöttes tények egyre inkább népszerűvé teszik a P2P hitelezést úgy a hitelfelvevők, mint a hitelezők számára.

A legelső P2P hitelezési platformot, a Zopa-t 2005-ben alapították az Egyesült Királyságban. A Zopát követően számos P2P hitelezési platform jött létre, köztük több mint 1000 Kínában. Az Amerikai Egyesült Államokban a Lending Club a legbefolyásosabb P2P hitelezési platform, amelyet 2014-ben bevezettek a New York-i tőzsdére. A Verified Market Research (2021) szerint a P2P hitelezés piacának értéke 2020-ban 84,89 milliárd USD volt és az előrejelzések szerint 2028-ra eléri az 578,03 milliárd dolláros értéket.

Az egyedi hitelkérelmek növekedésével a hitelkockázat értékelése egyre fontosabbá vált a pénzügyi szakemberek és kutatók számára is. Norbert (2022) szerint a hagyományos módszerek, amelyek sokáig dominálták ezt a területet, egyre inkább kihívásokkal szembesülnek a változó gazdasági környezet és az adatok bősége miatt. Az új technológiák, mint például a mesterséges intelligencia alapú algoritmusok, ígéretes alternatívát kínálnak a hagyományos modellek mellett.



A dolgozatban áttekintésre kerülnek az új megközelítések és ezek összehasonlítása a hagyományos módszerekkel hatékonyság és alkalmazhatóság szempontjából . Az elemzés célja, hogy betekintést nyújtson abba, hogy hogyan lehetnek az új technológiák hatékony eszközök a hitelminősítő intézmények számára a kockázatok kezelésében és a döntéshozatalban. A pontosság (accuracy) szempontjából hasonlítja össze a döntési fa, véletlen erdő, valamint a logisztikus regresszió teljesítményét. Ezáltal megmutatja, hogy melyik módszer nyújtja a legmegbízhatóbb eredményeket a hitelminősítés terén. A modellek teljesítményének elemzése vizsgálja, hogy az új technológiák mennyire képesek-e javítani a hitelminősítő intézmények hatékonyságát és eredményességét a hitelezési döntéshozatal folyamatában.

A tanulmány négy részből tevődik össze, az első fejezet a szakirodalmi áttekintést tartalmazza, ezt követi az adatok és a módszertan fejezet, amelyben az elemzésben használt adatbázis, ennek a tisztítási folyamata és a használt módszerek kerülnek bemutatásra. Ezt követi az eredmények elemzése, majd a következtetések levonása.



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Szakirodalmi áttekintés

A hagyományos banki hitelezés hosszú évek óta meghatározó szereplője a gazdaságnak, amely biztosítja a vállalkozásoknak a szükséges tőkét a működésükhöz. Ezzel szemben a P2P hitelezés újabb megközelítést jelent, amely során közvetlen kapcsolat alakul ki a kölcsönt igénylő és a befektetők között az online platformokon keresztül. Ezek a platformok alacsonyabb költségeket és gyorsabb folyamatokat kínálnak a hiteligénylőknek, miközben lehetőséget teremtenek azoknak a befektetőknek, akik hajlandóak közvetlenül részt venni a finanszírozásban. Az üzleti döntéshozóknak fontos megérteniük mindkét lehetőség előnyeit és korlátait annak érdekében, hogy hatékonyan tudjanak választani a finanszírozási lehetőségek között és maximalizálják vállalkozásuk növekedési potenciálját. A hitelezési döntés hasonló a két finanszírozási forrás esetén (Kgoroadira et al., 2023).

Berger et al. (2021) szerint a hitelminősítés egy folyamat, amelynek során a pénzügyi intézmények vagy hitelezők eldöntik, hogy egy adott hitelfelvevő (fizikai személy vagy vállalat) képes-e visszafizetni a kölcsönt a megállapított feltételek szerint. Ez a folyamat segít a hitelezőknek felmérni a kockázatot, amelyet a kölcsön nyújtása jelent, valamint segít döntést hozni arra vonatkozóan, hogy milyen feltételekkel adják a hitelt vagy, hogy egyáltalán megadják-e azt. Zhang et al. (2024) szerint a hitelminősítő pontszámok lehetővé teszik a hitelezők számára, hogy meghatározzák azt a valószínűséget, amellyel a hitelfelvevő fogja visszafizetni a kölcsönt a megállapított feltételek szerint. A hitelezők és a hitelfelvevők számára egyaránt fontos, hogy ezek a hitelminősítések megbízhatóak és pontosak legyenek.

A szakirodalomban számos tanulmány vizsgálja a hitelminősítő algoritmusokat és hasonlítja össze teljesítményüket (Bellotti és Crook, 2009; Dumitrescu et al., 2019; Fraisse és Laporte, 2022; Estran et al., 2022; Ahmed, 2023; Wang et al., 2023; Song et al., 2023). A hitelminősítő algoritmusok két nagy csoportját különíti el a szakirodalom: a hagyományos módszerek és a mesterséges intelligencia alapú hitelminősítő algoritmusok. A hagyományos hitelminősítő algoritmusok az 1970-es években kezdtek elterjedni (Ye et al., 2018). Ezek az algoritmusok általában statisztikai modelleken és számítógépes programokon alapulnak,



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

amelyek az ügyfelek pénzügyi adatait és egyéb tényezőket elemzik a hitelképességük meghatározására.

A mesterséges intelligencia alapú hitelminősítő algoritmusok viszont viszonylag új fejlesztések, amelyek az elmúlt 10 évben terjedtek el, 2010 óta (Zhou et al., 2019). Ezek az algoritmusok összetettebb adatelemzést végeznek és képesek nagyobb adatmennyiségeket és változatosabb adatforrásokat feldolgozni, mint például a szociális média aktivitás vagy az online vásárlási szokások. Fraisse és Laporte (2022) szerint a hagyományos módszerek általában statisztikai modellek vagy szabályalapú rendszerek, amelyek előre definiált szabályok vagy paraméterek alapján döntenek. Ezek a módszerek általában nem tudnak adaptálódni az új adatokhoz vagy változó környezeti feltételekhez. A leggyakrabban használt statisztikai módszerek a logisztikus regresszió és a Lasso-logisztikus regresszió.

A mesterséges intelligencia alapú algoritmusok képesek nagy mennyiségű adatot feldolgozni és összetett mintákat felfedezni a döntéshozatal során. A mesterséges intelligencia alapú modellek közé tartozik a döntési fa, a véletlen erdő (random forest), gradiens erősítés (XGBoost), valamint a tartóvektor gép algoritmus (Support Vector Machine, SVM).

Teply és Polena (2020) szerint a döntési fa egy strukturált modell, amely döntési szabályokat reprezentál fa alakú diagram formában. A döntési fa segítségével könnyen értelmezhető, hogy milyen változók alapján hozza meg a modell a döntéseit. Például a hitelminősítés során a döntési fa segíthet megállapítani, hogy milyen jövedelmi küszöbérték alatt vagy felett kerül egy ügyfél a teljesítő, vagy nemteljesítő kategóriába. A döntési fa az adatokat címkékkel különbözteti meg egy bizonyos kérdéssorozat szerint. Az ágakat bizonyos vágási kritériumok szerint generálja és minden egyes csomópont egy kategóriát, értéket vagy döntést képvisel. A döntési folyamat során az információ mennyisége fokozatosan növekszik, majd, amint nem tud tovább növekedni megáll és létrehoz egy végeredményt. Ez osztályozási fa-modellként használható, amely alkalmas diszkrét típusú adatok, például hitelminősítések előrejelzésére.



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Zhou et al. (2023) szerint a véletlen erdő több döntési fát alkalmaz az adatok osztályozására. A véletlen erdő hatékonyan kezeli a nagy számú változót és a változók közötti összefüggéseket, és ellenáll a túltanulásnak (overfitting), azaz általában jól általánosítható. Emellett rugalmasan kezeli a hiányzó adatokat és nem igényel sok előfeldolgozást.

Wang et al. (2022) szerint az XGBoost egy gépi tanulási algoritmus, ami az extra gradient boosting eljárás egyik változata. Ez az algoritmus hatékonyan kezeli a nagy mennyiségű adatot és kiváló teljesítményt nyújt a prediktív modellezésben, mint a hitelkockázat-értékelési modellt.

Bellotti és Crook (2009) szerint az SVM egy gépi tanulási módszer, mely a felügyelt tanulási problémák közé tartozik és a minták osztályozására vagy regresszióra használható. A hitelminősítésben az SVM-et használhatják arra, hogy becsléseket adjanak egy adott hitelkérelmet benyújtó ügyfél fizetési kockázatáról.

Dumitrescu et al. (2022) szerint a logisztikus regresszió valószínűségi alapú osztályozást alkalmaz, hogy előrejelezze, hogy egy ügyfél képes lesz-e törleszteni a hitelt vagy sem. A logisztikus regresszió használható azoknál a hitelezési platformoknál, amelyeknél a nemteljesítések azonosítási pontosság javítása a cél, valamint annak megállapítása, hogy a hitel késedelmes-e.

Fraisse és Laporte (2022) tanulmányozták a különbségeket a mesterséges intelligencia alapú algoritmusok és hagyományos módszerek között a hitelminősítés esetén. A módszerek közötti fő különbség a flexibilitás és az adatok elemzésének mélysége. A mesterséges intelligencia alapú algoritmusok, különösen a gépi tanulás és a mély tanulás technikái, sokkal rugalmasabbak és adaptívabbak, mint a hagyományos módszerek.

Dzik-Walczak és Heba (2021) összehasonlította a logisztikus regresszió és a döntési fa teljesítményét. A tanulmány egy felvetésen alapul, miszerint a legalacsonyabb téves osztályozást a döntési fa esetén, míg a legmagasabb helyes osztályozási arányt a logisztikus regresszió esetén érik el. Az elemzésben használt adatbázis a Lending Club oldaláról származik, 2011 és 2013 közötti periódusra vonatkozik és 119.160 hitelfelvevő adatát



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

tartalmazza. Megállapították, hogy a logisztikus regresszió jobb eredményt ért el, mint döntési fán alapuló hitelkockázati modell. Ugyancsak a logisztikus regresszió és a döntési fa teljesítményét hasonlította össze Jagtiani és Lemieux (2019), akik megerősítik az előbbi eredményt.

Dong és Ma (2018) a logisztikus regressziót hasonlította össze a véletlen erdővel. Az adatok a Lending Club 2019-es adatbázisából származnak és több mint 115.000 felhasználó hiteladatait tartalmazza. A véletlen erdő teljesített jobban az eredményeik alapján. Wu és Pan (2021) a logisztikus regressziót, valamint a véletlen erdőt alkalmazták az adatok korrelációjának optimalizálása és a 145 változót tartalmazó adatbázis 22 változóra való szűkítése érdekében. 2017 első negyedéből származnak az adatok, összesen 96.781 hitelfeltevőre (67.747 jó ügyfél és 29.034 rossz ügyfél) és 145 változóra vonatkoznak. A logisztikus regresszió bizonyult a legalkalmasabbnak az adathalmazon. A logisztikus regresszió és a véletlen erdő közötti összehasonlítást találjuk Woo és Sohn (2022), Chang et al. (2022) tanulmányában is, akik ugyanerre a következtetésre jutottak.

Wang és Chang (2022) tanulmányukban a magánszemélyek hitelkockázati jellemzőit vizsgálták logisztikus regresszióval, valamint XGBoost-on alapuló diszkriminációs modellel. Ezt követően a nemteljesítések közötti megkülönböztetések, a minél pontosabb eredmény eléréséhez, XGBoost-al történtek. A modell 10.000 valós hiteladatot használ. Összehasonlításra kerül a logisztikus regresszió, valamint az XGBoost. A logisztikus regresszió volt a legeredményesebb, mivel ennek volt a legmagasabb az AUC² értéke és a legnagyobb pontossággal rendelkezett. Arra a következtetésre jutottak, hogy az XGBoost-on alapuló hitelkockázat értékelési modell erős nemteljesítési megkülönböztető képességgel és robusztussággal rendelkezik.

² AUC (area under the curve) a modellek összehasonlításra használható. Az a modell hatékonyabb, megbízhatóbb, melynek nagyobb a ROC görbe alatti területe (AUC-ja).



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Zhou et al. (2019) tanulmánya egy véletlen erdő algoritmuson alapuló osztályozási módszert vizsgál, majd hasonlítja össze az eredményeit egy XGBoost algoritmussal. Ezt a hitelfelvevők hitelképességének osztályozásában alkalmazza egy P2P hitelezési platform adatain a nemteljesítési valószínűség előrejelzésére, valamint a hitelezési veszteség csökkentésére. A teljesítmény jelentős elfogadási szint alatt romlott. Tehát, ha a hitelező nagyon alacsony elfogadási szintet alkalmaz, vagyis nagyon kis kockázatot vállal, akkor a módszer hatékonysága csökken. Ez azt jelenti, hogy a módszer kevésbé lesz pontos vagy megbízható az alacsonyabb elfogadási szinteken. A költségérzékenység figyelembe vétele esetén sem sikerült jelentősen javítani a módszer nemteljesítő hitelfelvevők azonosítására vonatkozó képességét. A hagyományos modellek az egyéni szociodemográfiai információkat és a hitelkérelemmel kapcsolatos információkat bemeneti adatokként értelmezték és vizsgálták. A kérelmezők dinamikus tranzakciós előzményei azonban a viselkedésük fontos mérőszámai, de a hagyományos értékelési rendszerben nem szerepelnek. A tanulmányban szereplő adatok egy P2P hitelező cégtől származnak. Ez az adatkészlet 38 változót és 15.000 hitelfelvevő adatait tartalmazta. A modellben használt független változók az életkor, a családi állapot, valamint a családi összbevétel. Az eredmények azt mutatták, hogy a véletlen erdő algoritmus jobban teljesített, mint az XGBoost. Ágoston (2022) tanulmányban az SVM módszert hasonlították össze a neurális hálóval, valamint a véletlen erdővel. A csődkutatási módszerek közül a véletlen erdő teljesített a legjobban. Az AdaBoost³-ot, valamint a Bagging⁴-et használták az adatok kiegyensúlyozásához, elemzéséhez, valamint rendezéséhez. A csődmodell adatbázisán, ahol összesen 757 hitelt és 12 változót vettek figyelembe a neurális háló, valamint az SVM módszert elemezték. Az SVM több, mint 5%-kal magasabb

³ Az AdaBoost olyan gépi tanulási technika, amely, amely más modellekkel együttműködve alkalmazható az optimális teljesítmény eléréséhez, úgy, hogy gyengébb modelleket kombinál a pontosabb előrejelzések érdekében.

⁴ A bagging olyan együttes tanulási módszer, amely gyakran használatos a variancia csökkentésére, olyan adathalmazokban, amelyek hibákat vagy zavaró tényezőket tartalmaznak, ami megnehezíti vagy torzíthatja az adatok elemzését vagy értelmezését.



osztályozási pontosságot ért el ezen az adatbázison. Megállapították, hogy a csődelőrejelzésben lévő fejlődés leginkább a mesterséges intelligenciának, valamint a gépi tanulásnak köszönhető.



Adatok és módszertan

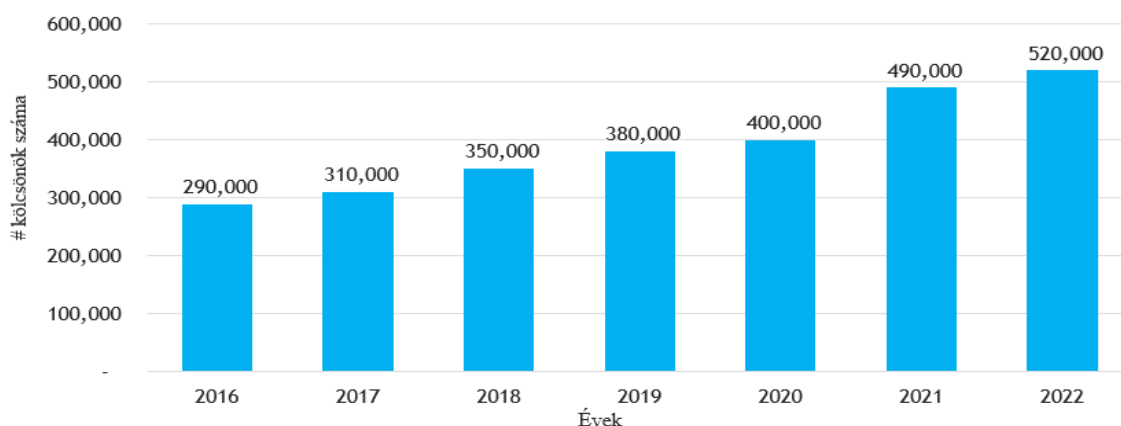
Ebben a fejezetben bemutatásra kerül az elemzésben használt adatbázis és az adattisztítási folyamat, amely magába foglalja a számomra nem releváns változók törlését, a változók transzformációját, a hiányzó és kiugró értékek kezelését és az alulmintavételezést. Az adatok bemutatása után az elemzésben használt módszerek rövid leírását is tartalmazza a fejezet.

Az adatok

Davaadorj et al. (2024) tanulmánya szerint a Lending Club egy olyan peer-to-peer hitelnnyújtó vállalat, amely egy online platformon keresztül hozza össze a hitelfelvevőket a befektetőkkel. Olyan embereknek nyújt hitelt, akiknek 1.000 és 40.000 dollár közötti személyi kölcsönre van szükségük. A hitelfelvevők a kölcsön teljes összegét megkapják, mínusz a kezelési költség, amelyet a vállalatnak fizetnek. A befektetők a személyi kölcsönökkel fedezett kötvényeket vásárolnak és a Lending Clubnak szolgáltatási díjat fizetnek. A vállalat megosztja a platformján keresztül bizonyos időszakokban kibocsátott összes hitel adatait.

Az alábbi ábrán szemléltetve van, hogy az elmúlt években hány új kölcsön került fel a Lending Club oldalára. Az adatokból is látszik a P2P hitelezés dinamikus fejlődése.

1. Ábra: A Lending Club által folyósított hitelek száma, 2016-2022



Forrás: saját szerkesztés Lending Klub adatai alapján.



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Az elemzés a Lending Club 2017-es első negyedévének hiteladataira épül. Az eredeti adatbázis 58.102 hitelfeltevő adatait tartalmazza. Az adatbázis 53.148 olyan ügyfelet tartalmaz, akik teljes mértékben visszafizetett a hitelt, illetve 4.954 nemteljesítő ügyfelet.

Az eredeti adatkészlet 117 változót tartalmazott, amelyből 35 változót választottam ki, köztük egy olyan változót, amely a végleges hitelstátuszt tartalmazza. A változók miatt kerültek törlésre, mert nem voltak relevánsak az elemzésem szempontjából.

A függő változó a nemteljesítés (loan_status), amelynek öt értéke volt, viszont számomra kettő volt releváns, emiatt a hasonló értékeket összevontam. Az 1. táblázatban láthatóak a Lending Club-ról származó hitelek aktuális státuszai, amelyek az kategóriákba sorolhatóak: teljesen visszafizetve (fully paid), leírásra került (charged off), nemteljesítő (default), folyamatban levő (current), türelmi időben levő (in grace period), 16-30 nap közötti késedelemmel rendelkezik (late 16-30 days), 31-120 nap közötti késedelemmel rendelkezik (late 31-120 days) A teljes mértékben visszafizetett hitelek kategóriába tartoznak a folyamatban levő, a türelmi idő, a késedelmes (16-30 nap, valamint a 31-120 nap) hitelek, továbbá a bank által jóváírt, de nemteljesítő hiteleket kategóriába tartoznak a nemteljesítő hitelek is. Az elemzés szempontjából csak a teljesen visszafizetett és a nemteljesítő hitelek relevánsak. A nemteljesítés változót átalakítottam bináris változóvá: a 0 a teljesen visszafizetett hiteleket jelöli, a nemteljesítő hiteleket 1-el jelölöm.

1. Táblázat: Hitelek száma az átalakítások előtt és után

Hitel státusza	Kölcsönök száma
<i>Előtte</i>	
Teljesen visszafizetve	17 806
Leírásra került	4 950
Nemteljesítő	4
Folyamatban levő	34 053
Türelmi időben lévő	314
16-30 nap közötti késedelemmel rendelkezik	154



31-120 nap közötti késedelemmel rendelkezik	821
Utána	
Teljesen visszafizetve	53 148
Leírásra került	4 954
Összesen	58 102

Forrás: A Lending Club oldaláról származó adatok alapján saját szerkesztés

A minősítés (grade) változó azt jelzi, hogy az ügyfelek milyen osztályozással rendelkeznek egy A-tól F-ig terjedő skálán. Az régiség (emp_length) változó a munkaviszony hosszát mutatja, amelynek értéke kevesebb mint 1 évtől 9 évig terjed. Az 1 évnél rövidebb munkaviszony azt jelenti, hogy a hitelkérelem benyújtása előtt kevesebb, mint 1 évet dolgozott a hiteligénylő az adott munkahelyen. A hitelfelvevőknek meg kellett jelölniük a hitel felhasználásának célját is.

Az alábbi táblázatban látható a változók neve, leírása, valamint, hogy közülük melyikek voltak átalakítva, és melyikek nem.

2. Táblázat: Lending Club-ról származó változók

Átalakítva	Változó neve	Leírása
Nem	loan_amnt	A hitel összege.
Igen	term	Hitelezési periódus hossza:36 vagy 60 hónap.
Igen	grade	Minősítési kategória.
Igen	empt_length	Munkaviszony hossza a jelenlegi munkahelyen.
Igen	home_ownership	A lakástulajdoni státusz, amelyet hitelfelvevéskor megadtak: bérlet, saját, lezálog, egyéb.
Igen	annual_income	A hitelfelvevő éves jövedelme.
Igen	loan_status	A hitel jelenlegi állapota: folyamatban, késedelmes, elfogadva, visszautasítva.
Igen	purpose	A hitel célja.
Nem	addr_state	Az állam, amiben laknak.
Nem	delinq_2yrs	A 30 napon túli késedelmes esetek száma az elmúlt 2 évből.
Nem	dti	A hitelfelvevő összes adósságkötelezettségének aránya a havi törlesztőrészletben.
Nem	inq_last_6mths	Megkeresések száma az elmúlt 1 hónapban.



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Átalakítva	Változó neve	Leírása
Nem	earliest_cr_line	A hitelfelvevő legkorábbi bejelentett hitelkerete megnyitásának napja.
Nem	open_acc	A számlák száma.
Nem	revol_util	A hitelfelvevő által felhasznált hitelösszeg az összes rendelkezésre álló rulírozó hitelhez viszonyítva.
Nem	total_acc	A hitelfelvevő hitelkereteinek száma.
Nem	chargeoff_within_12_months	A 12 hónapon belüli hitelkiesések száma.
Nem	delinq_amnt	Azoknak a számláknak a lejárt összege, amelyekkel a hitelfelvevő késik.
Nem	tax_liens	Adótartozások száma.
Nem	fico_range_low	A FICO pontszám legkisebb értéke.
Nem	fico_range_high	A FICO pontszám legnagyobb értéke.

Forrás: Kaggle-ről származó változók magyarázata

A táblázat az első oszlopában igen az átalakított váltókat jelenti. A term változó esetében karakter típusú volt, amelyeket az algoritmus hibának értelmezett. Ennek következtében szükséges volt faktor típusúvá átalakítani, így 41.684 darab 36 hónap időtartamú, valamint 16.418 darab 60 hónap időtartamú hitel maradt az adatbázisban. A minősítés (grade) változó is karakter típusú volt, amelyeket faktor típusúvá alakítottam át. Ennek következtében kaptunk 12.672 “A” osztályba sorolt, 17.173 “B” osztályba sorolt, 17.066 “C” osztályba sorolt, 7.344 “D” osztályba sorolt, 2.698 “E” osztályba sorolt, 804 “F” osztályba sorolt, valamint 345 “G” osztályba sorolt értéket. A régiség (emp_length) változó értékei is karakter típusúak, amelyek <1 -től, egészen 9-ig tartanak, faktor típusúvá alakítottam át. A lakástulajdon (home_ownership) változónak a következő értékei voltak: egyéb (any), jelzálog (mortgage), saját (own) és bérelt (rent). A 28.378 hitelfelvevőnek volt jelzáloggal terhelt lakása, 6.497 hitelfelvevő saját lakásban lakik, 23.219 bérelt lakásban, valamint 8 hitelfelvevő esetén az egyéb kategória volt megjelölve. Az egyéb kategóriát összevontam a bérelt kategóriával. A jövedelem (annual_inc) változó sok kiugró értékkel rendelkezett. Ennek a korrigálása érdekében meghatároztam egy felső határt, amely felé eső értékeket kiszűröm a változó adatai közül. Az adatbázisban megjelennek a fico_range_low, valamint a fico_range_high változók, amelyek a hitelgénylők legmagasabb és legalacsonyabb FICO pontszámait mutatják. Az értékek 600 és 700 közöttiek. A FICO Score egy általánosan



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

elfogadott hitelkockázat értékelő pontszám. A Fair Isaac Corporation (FICO) fejlesztette ki, és a hitelképesség értékelésére használják.

A változók szűrését, valamint átalakítását követően három módszer, a logisztikus regresszió, a döntési fa, valamint a véletlen erdő, teljesítményét hasonlítottam össze. Ezek közül a döntési fa és a véletlen erdő a mesterséges intelligencia alapú hitelminősítő algoritmusok közé tartoznak, míg a logisztikus regresszió a hagyományos módszerek legismertebb típusa. Jagtiani és Lemieux (2019) szerint több különbség van a módszerek között. A mesterséges intelligencia alapú módszerek általában képesek nagyobb és összetettebb adathalmazokkal dolgozni. Ezek az algoritmusok hatékonyan ki tudják használni az adatokban rejlő összefüggéseket és mintákat. Gyakran jobb teljesítményt nyújtanak a hagyományos módszereknél, mivel képesek bonyolultabb mintákat észlelni. Ez pontosabb és hatékonyabb eredményeket jelenthet a hitelminősítés során. A hagyományos módszerek sokkal könnyebben értelmezhetőek, valamint magyarázhatóak, mivel matematikai modellek alapján működnek.

A numerikus változók leíró statisztikáit a 3. táblázat tartalmazza. A jövedelem (annual_income) változó tartalmazott kiugró értékeket, a régiség (emp_length) változó rendelkezett hiányzó értékekkel.

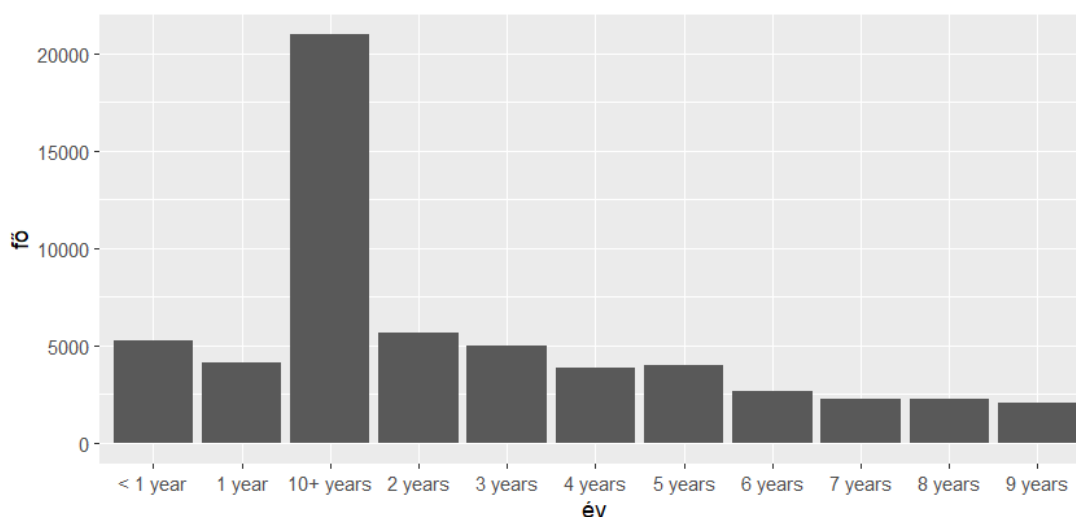
3. Táblázat: A numerikus változók leíró statisztikái

Változó	Min	Medián	Átlag	Max
loan_amnt	1.000,00	12.500,00	14.947,00	40.000,00
annual_income	32,00	65.000,00	72.168,00	170.520,00
delinq_2yrs	0,00	0,00	0,31	25,00
dti	0,00	18,29	19,39	999,00
inq_last_6mths	0,00	0,00	0,49	5,00
open_acc	1,00	11,00	11,7	81,00
revol_util	0,00	46,50	47,44	134,30
total_acc	2,00	21,00	23,35	141,00
fico_range_low	660,00	695,00	700,00	845,00
fico_range_high	664,00	699,00	704,00	850,00

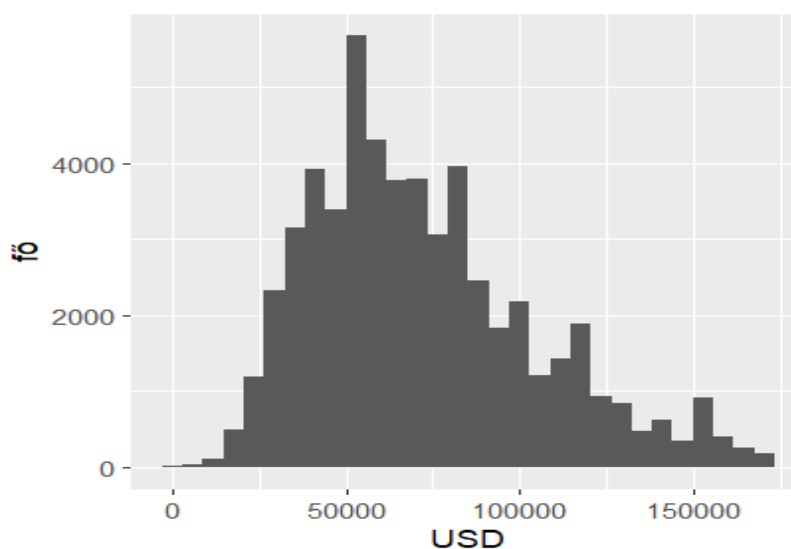
Forrás: saját számítás

A jövedelem (annual_income) változó kiugró értékeit vizsgáltam, a 170.900 USD-nál nagyobb jövedelmet tekintettem kiugró értéknek az interkvartilis terjedelem alapján. Az ügyfelek jövedelem, valamint régiség szerinti eloszlása az alábbi két ábrán látható. A legtöbb ügyfél 10 évnél nagyobb régiséggel rendelkezik. Az ügyfelek átlagos jövedelme 72.168 dollár.

2. Ábra: Az ügyfelek régiség szerinti eloszlása



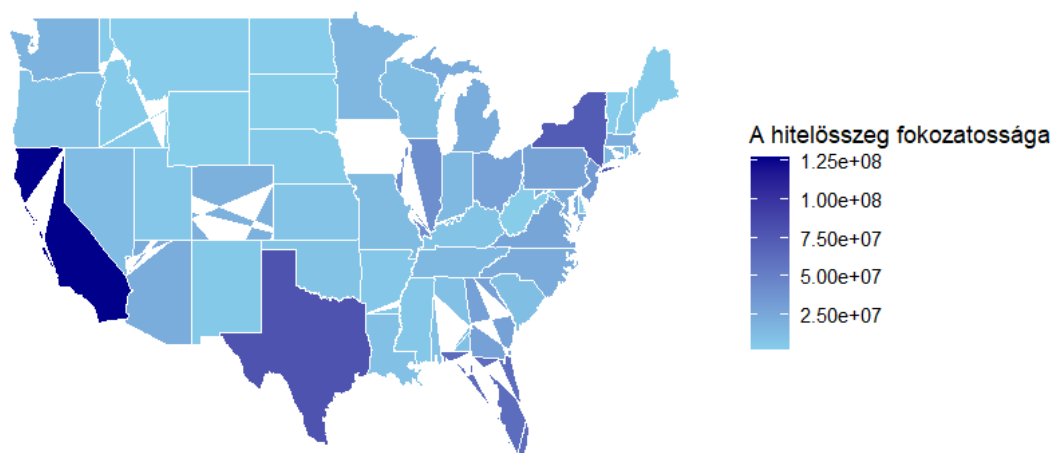
3. Ábra: Az ügyfelek jövedelem szerinti eloszlása



Forrás: saját szerkesztés

Az alábbi ábrán látható az átlagos hitelösszeg államok szerint hőtésképen ábrázolva. A fehér területek azt jelzik, hogy azokon a területeken egyáltalán nem vettek fel hitelt.

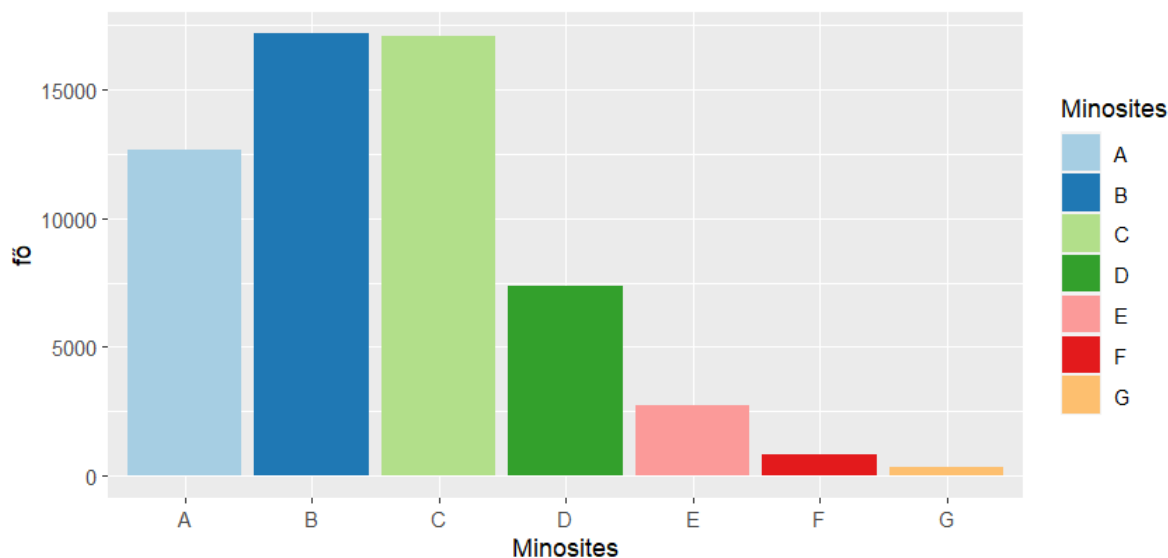
4. Ábra: A hitelösszeg átlaga államok szerint



Forrás: saját szerkesztés

A minősítés (grade) alapján hét osztályba vannak sorolva a hitelek. Az A kategóriájú hitel számít a legjobbnak, míg a G osztályban levő a legrosszabbnak, a hitelfelvevő adatai alapján vannak meghatározva a kategóriák. Az ügyfelek minősítés szerinti eloszlása az alábbi ábrán látható. A legtöbb ügyfél a B és a C minősítési kategóriába tartozik.

5. Ábra: Az ügyfelek minősítés szerinti eloszlása



Forrás: saját szerkesztés

Módszertan

A változók közötti kapcsolatokat t teszttel (a nemteljesítés és a numerikus változók esetén) és Chi-négyzet próbával (a nemteljesítés és a faktor típusú) tanulmányoztam.

Az ügyfelek nemteljesítési valószínűségének három módszer segítségével becslöm: logisztikus regresszió, döntési fa és véletlen erdő.

Logisztikus regresszió

Chang et al. (2022) szerint a legismertebb statisztikai módszer a nemteljesítési valószínűség becslésére a logisztikus regresszió. A függő változó bináris, 0 az értéke a jól teljesítő ügyfelek esetén, 1 pedig a nemteljesítő ügyfelek esetén. A binomiális logisztikus regresszió alakja a következő:

$$P(y = 1 | x_1, \dots, x_k) = T(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon),$$

ahol y egy binomiális függő változó, amely értéke 0 vagy 1. Továbbá az x -ek a független változók, amelyeket a T függvény leképez valós számokra 0 és 1 között, úgy, hogy a becslt



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

érték $[0,1]$ közötti legyen. Az ε a hibatagot jelöli. A következő egyenlet mutatja, hogy a T függvény hogyan képezi le ezeket az értékeket:

$$T(x) = \frac{e^x}{1 + e^x}$$

Ezt követően definiálásra került $\pi(x) = P(y = 1|x)$, amely a következő:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Ezt követően egy függvény maximalizálásával, amelyet log-likelihood-nak hívnak meghatározásra kerülnek az együtthatók. Brimacombe (2016) szerint ez a log-likelihood a bemeneti adatokra való illesztésének mértéke, valamint azt mutatja, mennyire valószínű, hogy a modell előrejelzi a megfigyelt adatokat.

A logisztikus regresszióban egy gyakran használt teszt a Chi-négyzet teszt, ami egy statisztikai eljárás, amelyet arra használható, hogy meghatározzuk, hogy van-e összefüggés két kategorikus változó között. A teszt összehasonlítja a megfigyelt és az elvárt gyakoriságokat. Ha a megfigyelt és az elvárt értékek jelentősen közelítenek egymáshoz, akkor megtartjuk a nullhipotézist, és arra következtethetünk, hogy összefüggés van a változók között.

Döntési fa

Chang (2022) szerint a döntési fa egy prediktív modell, amely egy fa struktúráját használja az adatok osztályozására vagy regressziójára. A döntési fa kezdetben egy gyökér csomópontból indul ki, amely tartalmazza az összes rendelkezésre álló adatot. Ezután az algoritmus kiválaszt egy attribútumot, amely felhasználva felosztja az adatokat különböző ágakra. A felosztások az attribútum értékétől függenek, és ezzel a fa növekszik, ahogy haladunk lefelé a gyökértől a levelekig. A döntési fa folytatja az attribútumok felosztását minden új ágon, és rekurzív módon halad lefelé a fa struktúrájában. Amikor elér egy leállási feltételt, például a fa maximális mélységét, az adatok egy bizonyos számú csoportra való felosztását vagy más előre meghatározott feltételeket, akkor leveleket hoz létre. Ezek a levelek tartalmazzák az osztályozás vagy regresszió eredményeit. A döntési fa előnyei közé tartozik az



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERENCEA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

egyszerűség és az érthetőség, mivel könnyen értelmezhető és magyarázható az eredménye. Azonban figyelembe kell venni, hogy hajlamos lehet az overfittingre, különösen bonyolult adathalmazokon. Emellett fontos az optimális paraméterezés és az overfitting elleni védelem alkalmazása annak érdekében, hogy hatékonyan működjön az adott probléma megoldása során.

Véletlen erdő

Teply és Polena (2020) szerint a véletlen erdő egy olyan ensemble tanulási módszer, amelyet eredetileg Ho (1995) javasolt, majd továbbfejlesztette egy bagging módszerrel Breiman (1996). Az algoritmus célja, hogy kezelje azt a problémát, hogy a döntési fák könnyen szabálytalan mintává válnak, minél több elágazásuk van. Ez a jelenség túlillesztéshez vezethet, mivel a variancia minden esetben magas lesz, még akkor is, ha alacsony az eltérés. A nagy számok törvénye alapján a véletlen erdők, olyan értéket konstruálnak, amelyek minimálisan eltérnek a döntési fáktól, hogy csökkentsék a variancia hatását. Ez a minimális eltérés a mintavételezésben nyilvánul meg. Általánosságban elmondható, hogy több döntési fával az előrejelzések eredményei pontosabbak lehetnek, ami egy jobb modellstabilitást jelent. Ez azonban hosszabb képzési időt is jelent, ezért a modell teljesítményének elérésekor figyelembe vesszük az egyensúlyt a képzési idő és a modell stabilitása között.



Eredmények

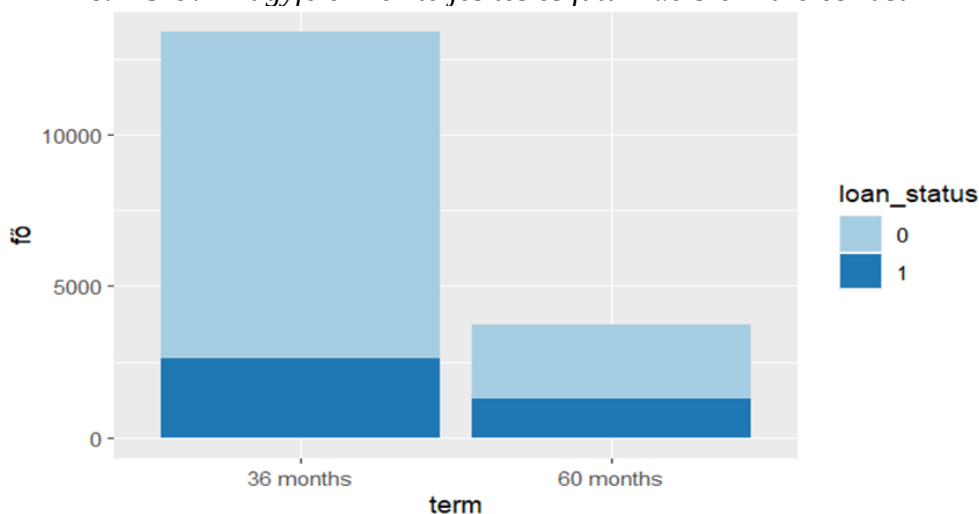
Az adattisztítás után az adatbázisban 13.182 olyan ügyfelet tartalmazott, akik teljes mértékben visszafizetették a hitelt, illetve 3.904 nemteljesítő (magas nemteljesítési kitettségű) ügyfél maradt.

3.1. Változók közötti kapcsolatok elemzése

A változók közötti összefüggéseket 2 módszerrel vizsgáltam, annak érdekében, hogy megállapítsam, hogy van-e köztük összefüggés. A nemteljesítés változó és a többi 34 változó közötti kapcsolatot vizsgáltam. A khi-négyzet próbát használtam a minőségi változók esetén és a t-teszt a mennyiségi változók esetén. A kapcsolatvizsgálat eredményeit az 5. táblázat tartalmazza.

A nemteljesítés és futamidő közötti kapcsolat a khi-négyzet próba segítségével vizsgáltam. A próba szignifikanciaszintje kisebb, mint 0,05, tehát van kapcsolat a két változó között, ami azt jelenti, hogy a nagyobb futamidejű változók esetén magasabb a nemteljesítő ügyfelek aránya, ahogy az alábbi ábrán is látható.

6. Ábra: Az ügyfelek nemteljesítés és futamidő szerinti eloszlása

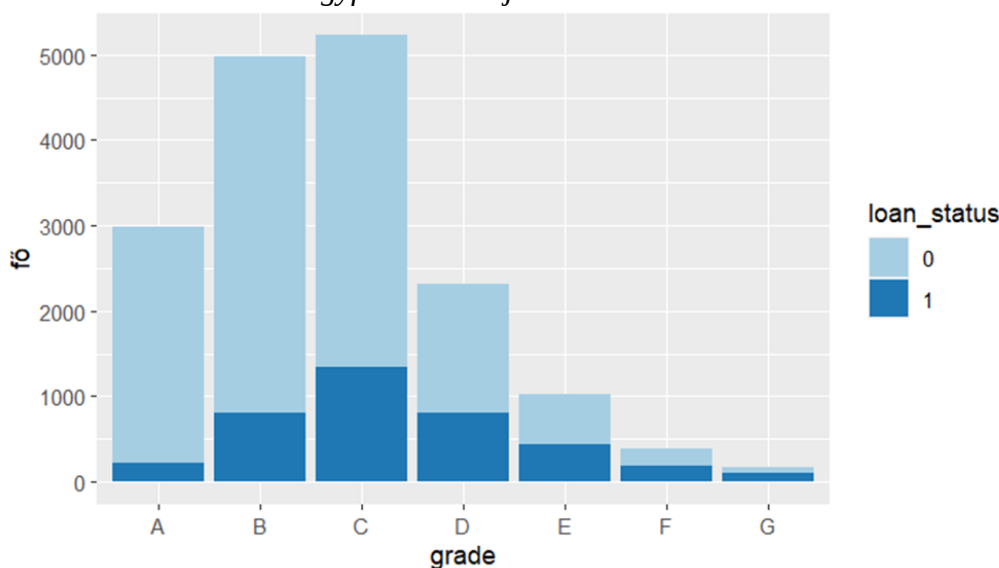


Forrás: saját szerkesztés



A nemteljesítés és a minősítés közötti kapcsolat is khi-négyzet teszttel vizsgáltam. Mivel a próba szignifikanciaszintje kisebb, mint 0,05, ezért van kapcsolat a nemteljesítés és a minősítés között. Minél rosszabb kategóriájú egy hitel, annál nagyobb a nemteljesítő ügyfelek aránya az adott kategóriában. A hitel minősítésének romlásával egyidejűleg csökken a hitelfelvevők száma is, ahogy azt az alábbi ábra mutatja.

7. Ábra: Az ügyfelek nemteljesítés és minősítés szerinti eloszlása



Forrás: saját szerkesztés

A nemteljesítés és az al-minősítés közötti kapcsolat is khi-négyzet próbával elemeztem. Van kapcsolat a nemteljesítés és az al-minősítés között. Minél rosszabb kategóriájú egy hitel, annál nagyobb a nemteljesítő ügyfelek aránya az adott kategóriában. A hitel minősítésének romlásával egyidejűleg csökken a hitelfelvevők száma is, ahogy azt az alábbi ábra mutatja.



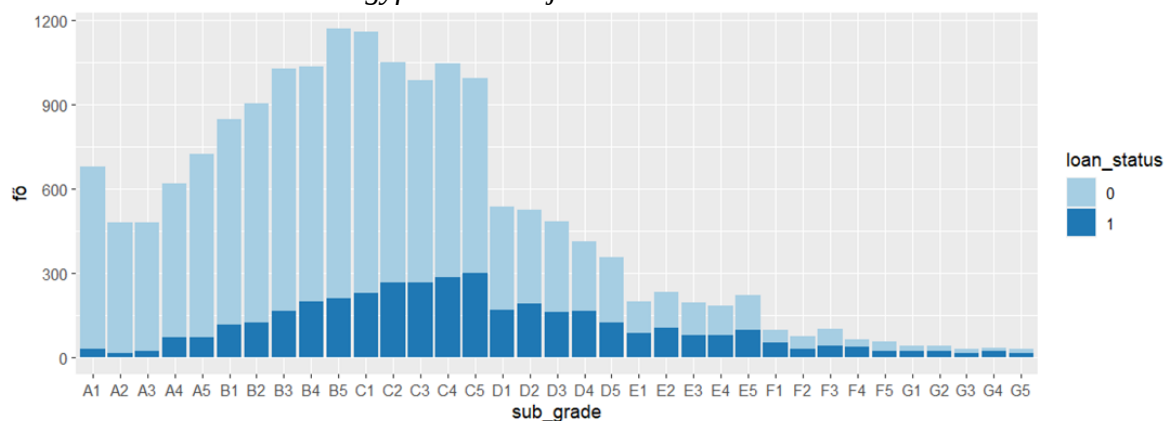
2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

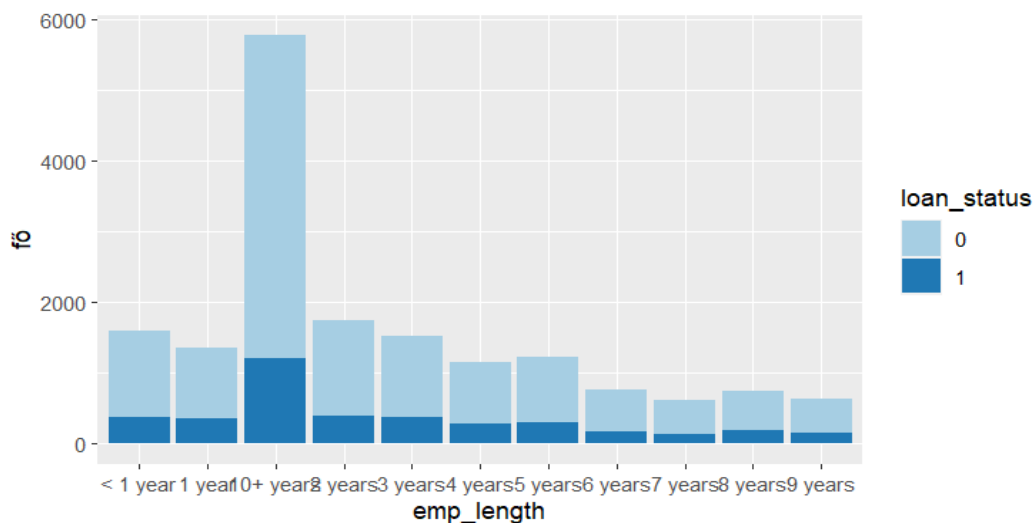
8. Ábra: Az ügyfelek nemteljesítés és az al-minősítés szerinti eloszlása



Forrás: saját szerkesztés

A nemteljesítés és a jelenlegi munkaviszony hossza közötti kapcsolat a khi-négyzet segítségével vizsgáltam. Az eredmények alapján van kapcsolat van a két változó között, ami azt jelenti, hogy minél hosszabb a jelenlegi munkaviszonya egy ügyfélnek, annál kisebb az esély a nemteljesítésre.

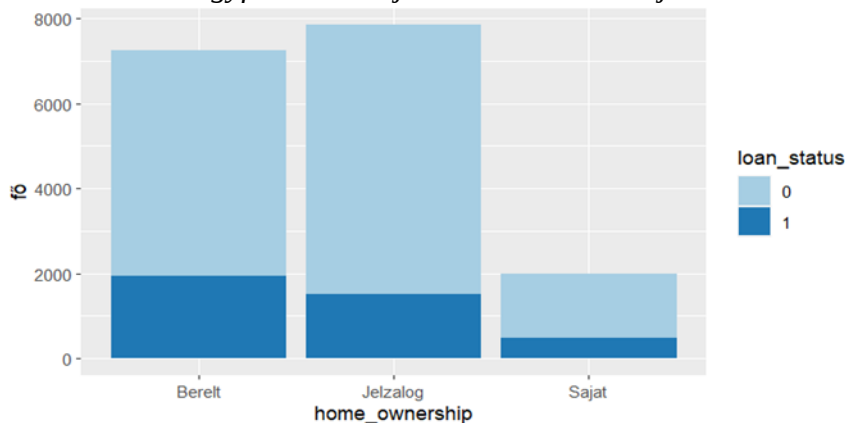
9. Ábra: Az ügyfelek nemteljesítés és a jelenlegi munkaviszony hossza szerinti eloszlása



Forrás: saját szerkesztés

A nemteljesítés és a lakástulajdon típusa közötti kapcsolat is khi-négyzet próbával tanulmányoztam és az eredmények alapján van kapcsolat nemteljesítés és a lakástulajdon között. A lakástulajdon lehet bérelt, jelzálog vagy saját. A bérelt kategória esetében a legnagyobb a nemteljesítő ügyfelek aránya az adott kategóriában.

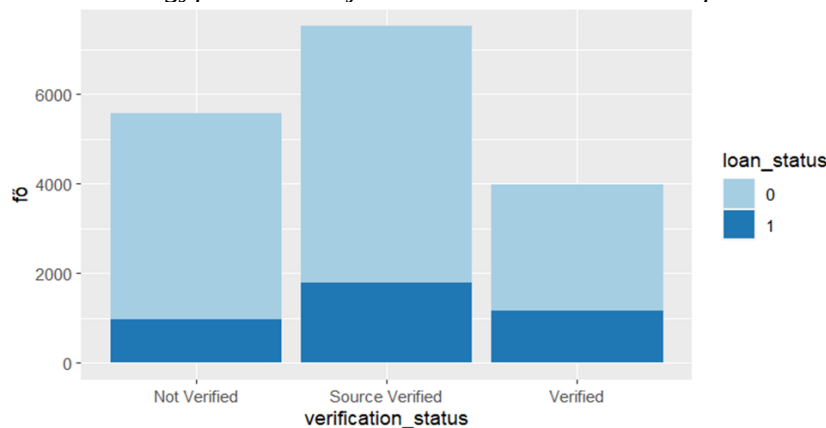
10. Ábra: Az ügyfelek nemteljesítés és a lakástulajdon szerinti eloszlása



Forrás: saját szerkesztés

A nemteljesítés a jövedelem ellenőrzési állapota közötti is van kapcsolat a khi-négyzet próba alapján. Jól látható, hogy az ellenőrzött források esetében, amely azt jelenti, hogy csak egy bizonyos része származik ellenőrzött forrásból, nagyobb a visszafizetés, és csak kevéssel nagyobb a nemteljesítés a nem ellenőrzöthöz, és teljesen ellenőrzöthöz képest.

11. Ábra: Az ügyfelek nemteljesítés és az ellenőrzési állapot szerinti eloszlása



Forrás: saját szerkesztés



2024. MÁJUS 15-18.

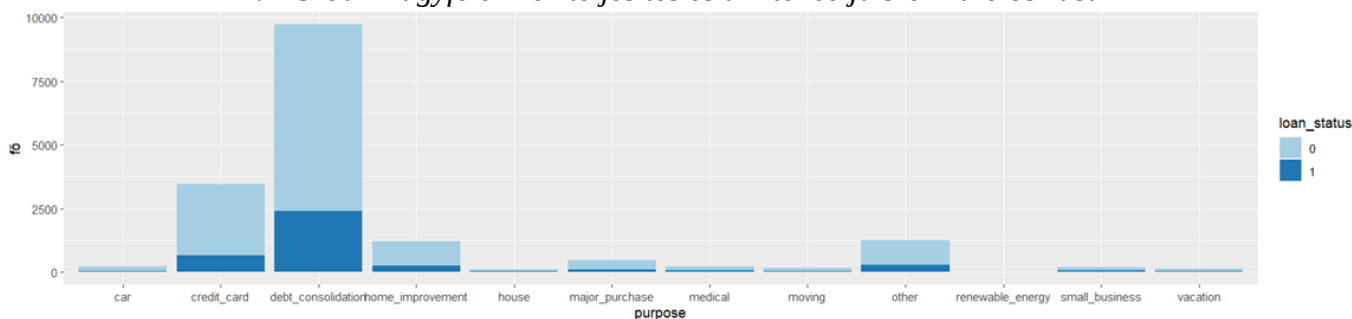
KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

A nemteljesítés és a hitel célja közötti kapcsolat is khi-négyzet próbával lett megvizsgálva. A próba szignifikanciaszintjét kisebb, mint 0,05, tehát van kapcsolat a nemteljesítés és a hitel célja között. Az adósság konszolidáció esetében volt a legnagyobb a visszafizetés, valamint a nemteljesítés is.

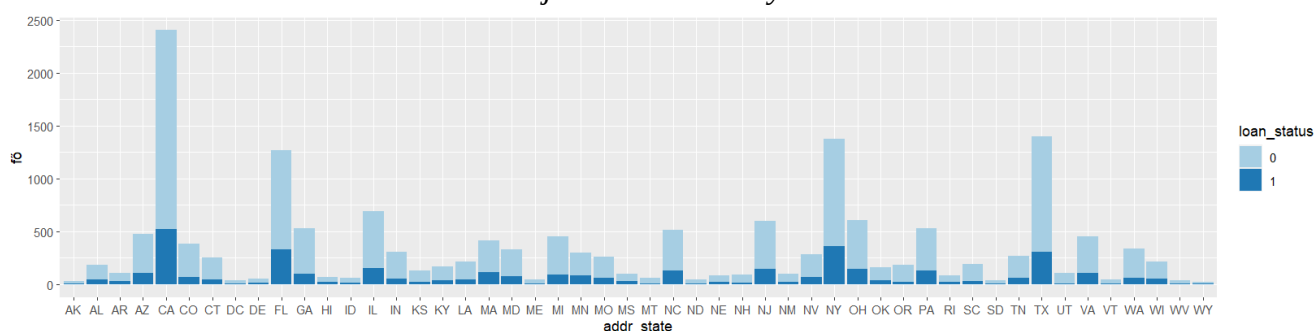
12. Ábra: Az ügyfelek nemteljesítés és a hitel célja szerinti eloszlása



Forrás: saját szerkesztés

A nemteljesítés és a lakhely közötti nincs kapcsolat is khi-négyzet próba eredményei alapján. Látható, hogy a legnagyobb nemteljesítés Kaliforniában volt, ahol a legtöbb hitelt is kérelmezték.

13. Ábra: Az nemteljesítés és a lakhely szerinti eloszlása



Forrás: saját szerkesztés

A mennyiségi változók és a nemteljesítés közötti kapcsolatot t-teszttel vizsgáltam, amelynek eredményei az alábbi táblázatban vannak összefoglalva. A táblázatban fel vannak tüntetve a változók nevei, a t statisztika értéke, a szabadságfok, valamint a szignifikanciaszint. Crook és Banasik (2012) szerint ez az érték méri, hogy mennyire különböznek a változók átlagai. Az abszolút érték minél nagyobb, annál nagyobb a különbség a csoportok átlaga között.



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

A 0,05 alatti szignifikanciaszint azt jelzi, hogy a mennyiségi változó várható értéke különbözik a nemteljesítő és a teljesítő ügyfelek esetén. A táblázat utolsó oszlopában fel van tüntetve, hogy a van-e kapcsolat a nemteljesítés és az adott mennyiségi változó között. Az “Igen” jelenti, hogy van kapcsolat, a “Nem”, pedig, hogy nincs, tehát nincs hatással a nemteljesítésre.

4. Táblázat: Mennyiségi változók és nemteljesítés közötti kapcsolat

Változó	t statisztika	szignifikanciaszint	Kapcsolat
loan_amnt	-12,013	< 0,001	Igen
annual_income	7,8311	< 0,001	Igen
dti	-12,76	< 0,001	Igen
fico_range_low	22,491	< 0,001	Igen
fico_range_high	22,491	< 0,001	Igen
revol_util	-12,148	< 0,001	Igen
total_acc	2,3481	< 0,018	Igen
total_rev_hi_lim	9,6972	< 0,001	Igen
bc_open_to_buy	15,852	< 0,001	Igen
mo_sin_rcnt_rev_tl_op	8,0718	< 0,001	Igen
total_cu_tl	-11,336	< 0,001	Igen
pub_rec_bankruptcies	-2,9122	0,003	Igen
delinq_2yrs	-3,0316	0,002	Igen
inq_last_6mths	-10,203	< 0,001	Igen
open_acc	-3,0921	< 0,001	Igen
pub_rec	-1,3213	0,186	Nem
acc_now_delinq	-1,0645	0,287	Nem
total_cu_tl	1,5701	0,116	Nem
chargeoff_within_12_mt	0.74317	0,457	Nem
hs			
num_actv_rev_tl	-12,329	< 0,001	Igen
num_il_tl	0,21468	0,831	Nem
num_op_rev_tl	-4,5919	< 0,001	Igen
num_tl_op_past_12m	-8,7989	< 0,001	Igen

Forrás: saját számítás

A nemteljesítést a következő változók befolyásolják a t-teszt alapján: a loan_amnt (hitel összege), az annual_income (jövedelem), a dti (adósság arány havi törlesztéssel), a fico_range_low (legkisebb FICO érték), a fico_range_high (legnagyobb FICO érték), a



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERENCEA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

revol_util (rulírozó hitelhez felhasznált hitelösszeg arány), a total_acc (hitelkeret száma), a total_rev_hi_lim (hitelkeret felső határa), a bc_open_to_buy (hitelkeret aránya), a total_cu_tl (hitelszámla számok), a mo_sin_rcnt_rev_tl_op (rulírozó hitelkeret megnyitás napja), a pub_rec_bankruptcies (csődök száma), a inq_last_6mths (megkeresések száma 6 hónapra), a delinq_2yrs (30 nap késedelmek), az open_acc (számlák száma), a num_actv_rev_tl (aktív rulírozó hitelek), a num_op_rev_tl (megnyitott rulírozó hitelek), valamint a num_tl_op_past_12m (egy éven belül megnyitott hitelszámlák).

3.2. Változók kategorizálása és a WoE érték kiszámítása

Először is, a woebin függvény segítségével létrehoztam egy binninget az adatok adathalmazból. A következő lépésben az információs értéket (IV) számítom ki az adatok adathalmaz minden változójára a nemteljesítés változó szerint. Az IV értékek segítségével értékeltem a változók fontosságát a modell szempontjából, ahol nagyobb IV értékek jelzik a változók fontosságát, minél nagyobb prediktív előrejelzését. A WoE érték a logisztikus regressziónál merült fel. Megvizsgáltam a logisztikus regressziót a Klasszifikációs mátrixal, valamint a ROC görbét is megnéztem, ezután létrehoztam a Logisztikus Regressziót WoE transzformált adatokat, majd itt is megvizsgáltam a Klasszifikációs Mátrixot, valamint a ROC görbét. A Logisztikus Regresszió WoE transzformált adatokkal jobb eredményeket mutatott, tehát pontosabban szűri ki a hibákat, mint a Logisztikus Regresszió.

3.3 Tanuló és teszt adatbázis

Az adatbázist felosztottam tanuló és teszt adatbázisra. A felosztást követően az eredeti adatbázis 70%-a lett a tanuló adatbázis, valamint 30%-a lett a teszt adatbázis. A tanuló adatbázisban 11.961 adat van, míg a teszt adatbázisba 5.125 adat tartozik. A tanuló adatbázison történik a modellek becslése, a teljesítményük értékelése pedig a teszt adatbázison. A tanuló adatbázis 9.228 jó ügyfelet, valamint 2.733 rossz ügyfelet tartalmaz, a teszt adatbázis 3.954 jó ügyfelet, valamint 1.171 rossz ügyfelet tartalmaz.



3.3.1. Logisztikus regresszió becslése

Első lépésben Stepwise logisztikus regressziót becsültem a nemteljesítési valószínűség becslésére a tanuló adatbázison. Estran et al. (2022) szerint ez egy eljárás a logisztikus regresszió modellépítés során, amely automatikusan kiválasztja és hozzáadja a változókat a modellhez, vagy eltávolítja őket a modellből a prediktív teljesítmény optimalizálás eredményeképp. Kezdetben a modellbe választott változókat kiválasztjuk a tanuló adatbázis alapján. A stepwise logisztikus regresszió lépésről lépésre döntést hoz arról, hogy mely változókat kell hozzáadni vagy eltávolítani a modellből. Tehát használata lehetővé teszi a modell automatikus optimalizálását és egyszerűsítését anélkül, hogy a manuálisan döntést kellene hozni minden egyes változóról.

A minősítés (grade), a jelenleg használt aktív rulírozó hitelszámlák száma (num_actv_rev_tl), a lakástulajdon (home_ownership), a hitel hossza (term), az összes rendelkezésre álló hitelkeret felső határa (total_rev_hi_lim), a hitel összege (loan_amnt), az éves bevétel (annual_inc), a FICO pontszám legkisebb értéke (fico_range_low), a megkeresések száma (inq_last_6mths), az összes adósságkötelezettség aránya (dti), a jövedelem ellenőrzési állapota (verification_status), az adótartozások száma (tax_liens), valamint a hitel célja (purpose) változók bizonyultak szignifikánsnak. A logisztikus regresszió eredményeit a következő táblázat tartalmazza.

5. Táblázat: *A nemteljesítési valószínűséget befolyásoló tényezők a logisztikus regresszió alapján*

Változó	Együttható	Std. hiba	Z érték	Pr(> z)	Esélyhányados
Konstans	1,458	0,804	1,813	0,070	-
gradeB	0,456	8,057	4,570	< 0,001*** ⁵	1,578

⁵Az együtthatók mellett található csillagok (*, **, ***) jelzik a szignifikanciaszintet "****" nagyon szignifikáns ($p < 0,001$), a "***" szignifikáns ($p < 0,01$), valamint a "*" szignifikáns ($p < 0,05$), . szignifikáns ($p < 0,1$), míg a hiányzó csillagok esetén nincs szignifikancia ($p \geq 0,1$). Minél inkább szignifikáns egy változó, annál jobban befolyásolja a modellt.



2024. MÁJUS 15-18. KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA

REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Változó	Együttható	Std. hiba	Z érték	Pr(> z)	Esélyhányados
gradeC	0,765	0,101	7,541	< 0,001 ***	2,149
gradeD	1,036	0,112	9,231	< 0,001 ***	2,818
gradeE	1,214	0,130	9,329	< 0,001 ***	3,367
gradeF	1,275	0,170	7,490	< 0,001 ***	3,579
gradeG	1,440	0,218	6,591	< 0,001 ***	4,220
num_actv_rev_ tl	0,081	0,009	8,470	< 0,001 ***	1,084
home_ownershi pJelzalog	-0,405	0,052	-7,677	< 0,001 ***	0,667
term 60 months	0,436	0,061	7,087	< 0,001 ***	1,546
total_rev_hi_li m	-0,004	< 0,001	-5,424	< 0,001 ***	0,996
loan_amnt	< 0,001	< 0,001	7.705	< 0,001 ***	1,000
annual_inc	-0,046	< 0,001	-4,727	< 0,001 ***	0,955
fico_range_low	-0,005	0,001	-5,664	< 0,001 ***	0,995
inq_last_6mths	0,107	0,026	4,045	< 0,001 ***	1,113
dti	0,010	0,003	3,452	< 0,001 ***	1,010
verification_sta tusSource Verified	0,149	0,057	2,606	0,009 **	1,161
verification_sta tusVerified	0,162	0,065	2,468	0,014 *	1,176
tax_liens	-0,153	0,072	-2,102	0,036 *	0,858
purposeSmall_ business	1,063	0,319	3,322	0,001 ***	2,894
purposeVacatio n	0,950	0,343	2,769	0,006 **	2,586

Forrás: saját számítás

Az eredmények szerint az egyes minősítés kategóriák kiemelkedő fontossággal rendelkeznek a nemteljesítés előrejelzésében. A nemteljesítési valószínűséget az előző táblázatban feltüntetett változók befolyásolják.

A teszt adatbázison teszteltem a logisztikus regresszió előrejelző képességét. Az osztályozási hatékonyságot a klasszifikációs mátrix alapján vizsgáltam. A 7. táblázat a



Az AUC vagyis a görbe alatti terület 0,6318, tehát a modell egy közepesen jó osztályozásba sorolható.

Ezután a Woe transzformált adatokon is becsültem a logisztikus regressziót. Itt a Woe adatbázist osztottam fel tanuló és teszt adatbázisokra. Ugyanúgy 70% és 30% volt a felosztási arány. Stepwise logisztikus regressziót használva megállapítottam, hogy a hitel összege (loan_amnt_woe), a futamidő (term_woe), az al-minősítés (sub_grade_woe), a munkaviszony hossza (emp_length_woe), a lakástulajdon (home_ownership_woe), az éves bevétel (annual_inc_woe), a jövedelem ellenőrzési állapota (verification_status_woe), a lakhely (addr_state_woe), az összes adóssághkötelezettség aránya (dti_woe), a legkorábbi bejelentett hitelkeret (earliest_cr_line_woe), a FICO pontszám legkisebb értéke (fico_range_low_woe), a megkeresések száma (inq_last_6mths_woe), a hitelkeretek száma (total_acc_woe), az összes rendelkezésre álló hitelkeret felső határa (total_rev_hi_lim_woe), valamint a hitelfelvevő által jelenleg használt aktív rulírozó hitelszámlák száma (num_actv_rev_tl_woe) bizonyultak szignifikánsnak. Az eredmények az alábbi táblázatban láthatóak.

7. Táblázat: *A nemteljesítési valószínűséget befolyásoló tényezők WOE transzformációval*

Változó	Együttható	Std. hiba	Z érték	Pr(> z)
Konstans	-1,217	0,023	-51,574	< 0,001 ***
loan_amnt_woe	0,908	0,115	7,890	< 0,001 ***
term_woe	0,439	0,080	5,461	< 0,001 ***
sub_grade_woe	0,469	0,087	5,366	< 0,001 ***
emp_length_woe	0,489	0,248	1,969	< 0,001 *
home_ownership_woe	0,943	0,122	7,713	< 0,001 ***
annual_inc_woe	0,747	0,169	4,421	< 0,001 ***
verification_status_woe	0,261	0,094	2,770	< 0,001 **
addr_state_woe	0,879	0,238	3,681	< 0,001 ***
dti_woe	0,467	0,110	4,251	< 0,001 ***
earliest_cr_line_woe	0,649	0,222	2,914	< 0,001 **
fico_range_low_woe	0,203	0,086	2,341	< 0,001 *
inq_last_6mths_woe	0,564	0,123	4,568	< 0,001 ***
total_acc_woe	1,142	0,293	3,895	< 0,001 ***

total_rev_hi_lim_woe	0,763	0,215	3,537	< 0,001 ***
num_actv_rev_tl_woe	0,697	0,148	4,701	< 0,001 ***

Forrás: saját számítás

Az alábbi táblázatban logisztikus regresszió teljesítménye a klasszifikációs mátrix alapján.

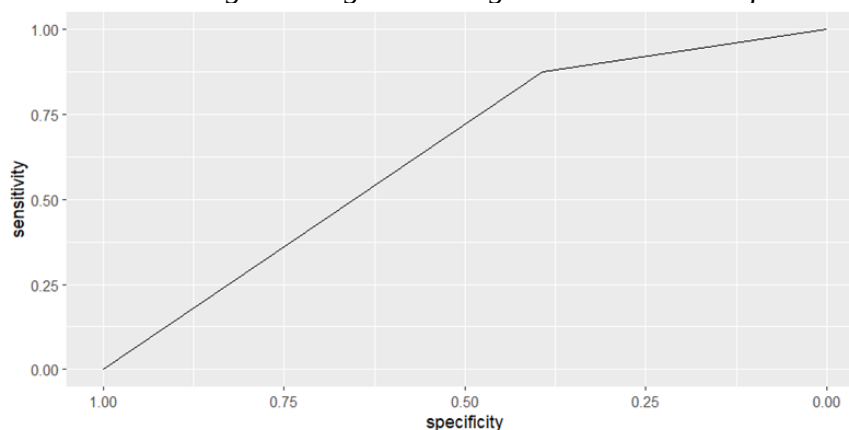
8. Táblázat: Klasszifikációs mátrix: logisztikus regresszió WOE transzformációval

	Tényleges	
Előrejelzett	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	1 552	147
Nemteljesítés (1)	2 402	1 024

Forrás: saját számítás

Az elemzés alapján a pontosság mutató (Accuracy) értéke 0,5026, ami azt jelenti, hogy 50,26%-ban helyesen osztályozta az ügyfeleket a logisztikus regresszió Woe transzformációval. Az érzékenység mutató (Sensitivity) 0,3925, ami azt mutatja, hogy a pozitív osztályba tartozó megfigyelések 39,25%-a helyesen lett azonosítva. A specificitás (Specificity) mutató értéke 0,8745, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések csak 87,45%-a lett helyesen azonosítva. A pozitív előrejelzési érték (Positive Predictive Value) 91,35%, míg a negatív előrejelzési érték (Negative Predictive Value) 29,89%.

15. Ábra: ROC görbe: logisztikus regresszió WoE transzformációval



Forrás: saját szerkesztés



Az AUC vagyis a görbe alatti terület 0,6335, tehát a modell közepesen jól megkülönbözteti a teljesítő és nemteljesítő ügyfeleket az adott adathalmazon, mint a nem transzformált adatokon becsült logisztikus regresszió.

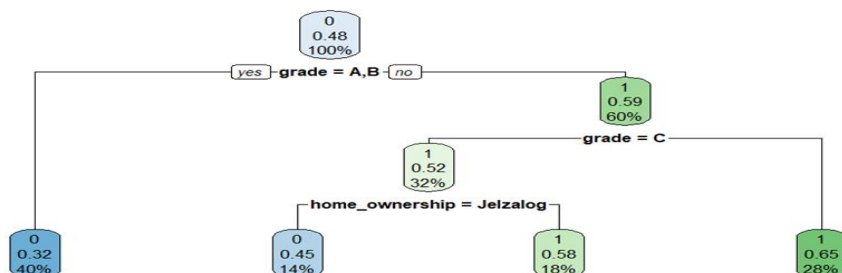
3.3.2. Döntési fa

A döntési fát elsősorban alulmintázással használtam. Moscato et al. (2021) szerint az alulmintázás kiegyensúlyozottabbá teszi az adatokat, ami segíti abban, hogy ne csak a többségi osztályra optimalizáljon. Ez javítja a modell pontosságát, az osztályozó helyes előrejelzéseinek aránya, sokkal pontosabb lesz. Az alulmintázás csökkenti az adatkészlet méretét, ami gyorsabb és hatékonyabb modellt eredményez. Létrehoztam az alulmintázott tanuló adatbázist, amelyben 2 951 jó ügyfél, valamint 2 733 rossz ügyfél volt.

Továbbá meghatározásra került az X-relative error értéke is, amely megmondja, hogy a keresztvalidáció során mért elemzés szerint, a fa mennyire teljesít jól az adott adathalmazon. Az érték minél kisebb, annál jobban általánosítható a kapott döntési fa. Az érték ebben az esetben 3 lett, amely azt jelenti, hogy a harmadik sorban van.

Ezt követően a kapott értékkel metszettem a döntési fát, eredményképpen azt kaptam, hogy amennyiben a hitel A vagy B minősítésbe lett besorolva, a nemteljesítés valószínűsége 0,32, tehát 32% volt. A C minősítésű hitelek esetén a következő változó, amit figyelembe vesz a lakástulajdon. A CP értékkel metszett döntési fa a következő ábrán figyelhető meg.

16. Ábra: Metszett alulmintázott döntési fa



Forrás: saját szerkesztés



A teszt adatbázison teszteltem az alulmintázott döntési fa előrejelző képességét. Ennek az elemzésnek a vizsgálatára a klasszifikációs mátrixot használtam. Az eredmények az alábbi táblázatban találhatók.

9. Táblázat: Klasszifikációs mátrix: alulmintázott döntési fa

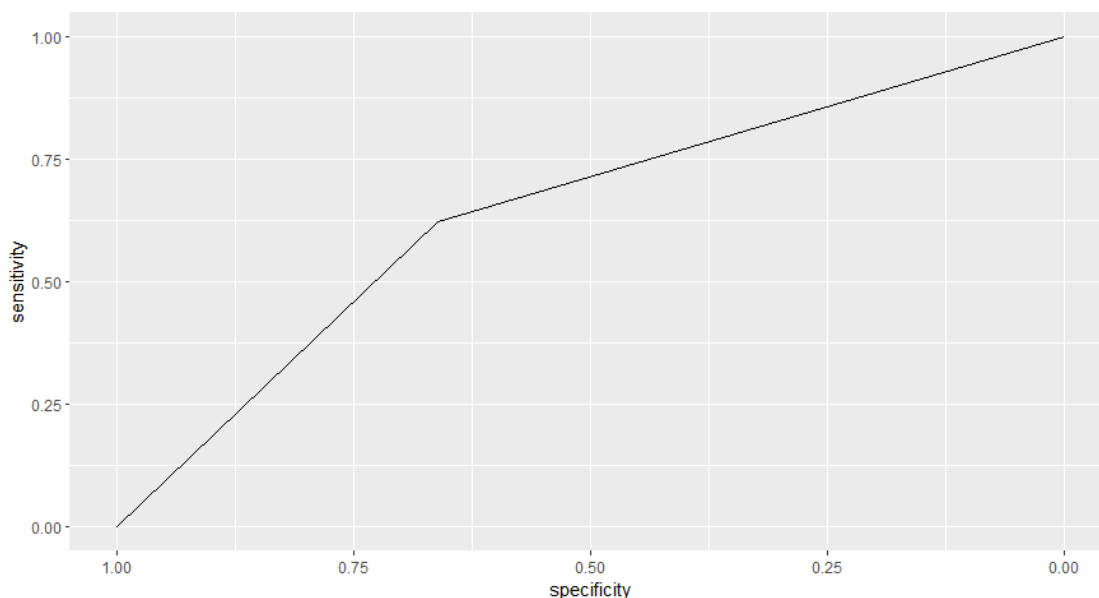
	Tényleges	
Előrejelzett	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	2 615	442
Nemteljesítés (1)	1 339	729

Forrás: saját szerkesztés

Az elemzés alapján a pontosság mutató (Accuracy) értéke 0,6525, ami azt jelenti, hogy 65,25%-ban helyesen osztályozta az ügyfeleket az alulmintavételezett döntési fa. Az érzékenység mutató (Sensitivity) 0,6225, ami azt mutatja, hogy a pozitív osztályba tartozó megfigyelések 62,25%-a helyesen lett azonosítva. A specificitás (Specificity) mutató értéke 0,6614, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések csak 66,14%-a lett helyesen azonosítva. A pozitív előrejelzési érték (Positive Predictive Value) 35,25%, míg a negatív előrejelzési érték (Negative Predictive Value) 85,54%.

A klasszifikációs mátrix után megrajoltam a ROC görbét, ami a következő ábrán látható.

17. Ábra: ROC görbe: döntési fa alulmintázással



Forrás: saját szerkesztés

Az AUC, vagyis a görbe alatti terület értéke 0,642, tehát a modell egy közepesen jó osztályozásba sorolható.

Shi et al. (2019) szerint a veszteség mátrix egy olyan technika, amely figyelembe veszi az osztályozási hibák költségeit a modell építése során. Ez hasznos egyensúlyhiányos adatkészletek esetén, ahol az egyes osztályok előrejelzési hibái különböző mértékű költségekkel járhatnak. Három különböző költség értéket határoztam meg, amennyiben helyes a besorolás a döntési fán belül, a költség 0, amennyiben egy rossz hitelről van szó, ami jó besorolást kapott, a költsége 10 lesz. A harmadik esetben, amennyiben van egy jó hitel, viszont ez rosszként van besorolva, a költség egyenlő lesz 1-el. A gyökércsomópontban, az adathalmaz kiinduló pontjában a rossz hitel aránya 0,77151 volt, tehát 77,15%-os lett. Az alábbi döntési fánál a komplexitási paraméter szintén 0,001-re volt állítva.

A döntési fa esetében a minősítés (grade), a rendelkezésre álló, fel nem használt hitelkeret összegének aránya a bankkártyák esetében (bc_open_to_buy), az összes rendelkezésre álló hitelkeret felső határa (total_rev_hi_lim), a munkaviszony hossza



(`emp_length`), a FICO pontszám legkisebb értéke (`fico_range_low`), valamint a hitelfelvevő által felhasznált hitelösszeg az összes rendelkezésre álló rulírozó hitelhez viszonyítva (`revol_util`) változók alsó határai a legfontosabbak a hitel minősítésének meghatározásában.

A veszteségmátrix segítségével meghatározott döntési fát előrejelzési képességének vizsgálatához a klasszifikációs mátrixot használtam. A következő táblázatban láthatóak az eredmények.

10. Táblázat: *Klasszifikációs mátrix: döntési fa a veszteségmátrix bevonásával*

Előrejelzett	Tényleges	
	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	2 615	442
Nemteljesítés (1)	1 339	729

Forrás: saját szerkesztés

Az elemzés alapján a pontosság mutató (Accuracy) értéke 0,6525, ami azt jelenti, hogy 65,25%-ban helyesen osztályozta az ügyfeleket a döntési fa és veszteség mátrix. Az érzékenység mutató (Sensitivity) 0,6225, ami azt mutatja, hogy a pozitív osztályba tartozó megfigyelések 62,25%-a helyesen lett azonosítva. A specificitás (Specificity) mutató értéke 0,6614, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések csak 66,14%-a lett helyesen azonosítva. A pozitív előrejelzési érték (Positive Predictive Value) 35,25%, míg a negatív előrejelzési érték (Negative Predictive Value) 85,54%.

Miután meghatároztam a döntési fa és veszteség mátrix Konfúziós mátrixát, a pontosabb és precízebb összehasonlítás érdekében, ábrázoltam a ROC görbét is, amely a következőképpen alakult.



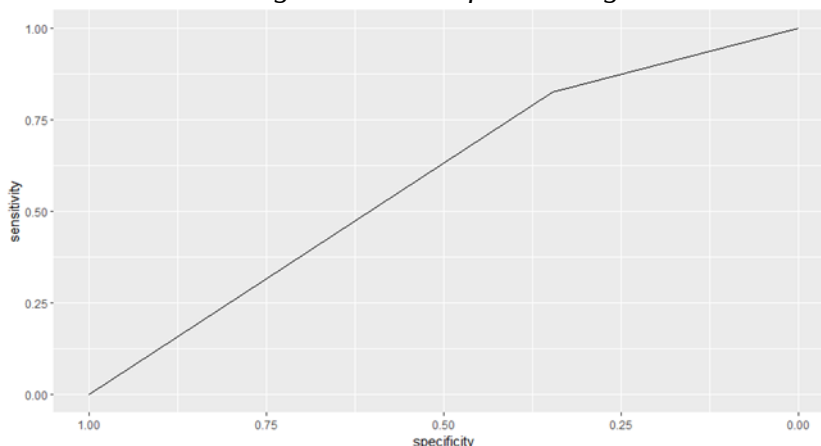
2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

18. Ábra: ROC görbe: döntési fa veszteségmátrix bevonásával



Forrás: saját szerkesztés

Az AUC vagyis a görbe alatti terület 0,5866, tehát a modell egy gyenge osztályozásba sorolható. Tehát a veszteségmátrixnál jobb eredményt mutatott az alulmintavételezés.

Ezek után alkalmaztam a döntési fát előzetes valószínűségek megadásával. Minimális módosítást végeztem az adatokon, tehát újra csoportosítottam őket, más százalékos felosztásban, tehát 60%-a az adatoknak tartozott a nemteljesítők közé, valamint 40% volt a teljesítőknél. A komplexitási paraméteren a döntési fa esetében 0,001 volt. Az előzetes valószínűségek megváltoztatásával becsült metszett döntési fa a következő ábrán figyelhető meg. Az A vagy B osztályozási valószínűsége 0,25 volt, tehát 25%. Továbbá az osztályozási valószínűsége a D, E, F, G minősítésű hiteleknek 0,53 volt, ami 53%-os, abban az esetben, hogyha a hitel összege (loan_amnt), kisebb mint 8 988, valamint az éves bevétel (annual_inc) kisebb, mint 68 500.



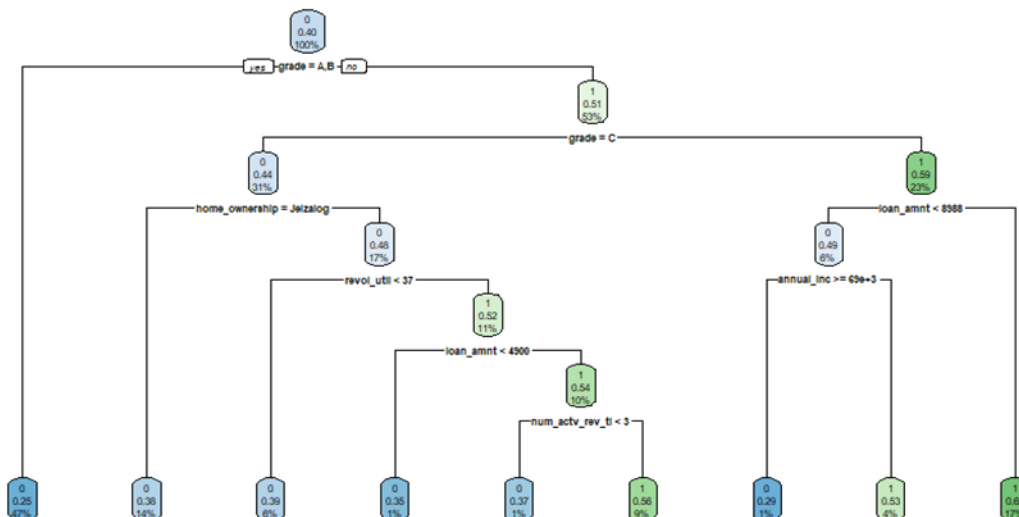
2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

19. Ábra: *Döntési fa előzetes valószínűségek megváltoztatásával*



Forrás: saját szerkesztés

Az alábbi táblázatban látható a döntési fa esetén a klasszifikációs mátrix.

11. Táblázat: *Klasszifikációs mátrix: döntési fa előzetes valószínűségek megváltoztatásával*

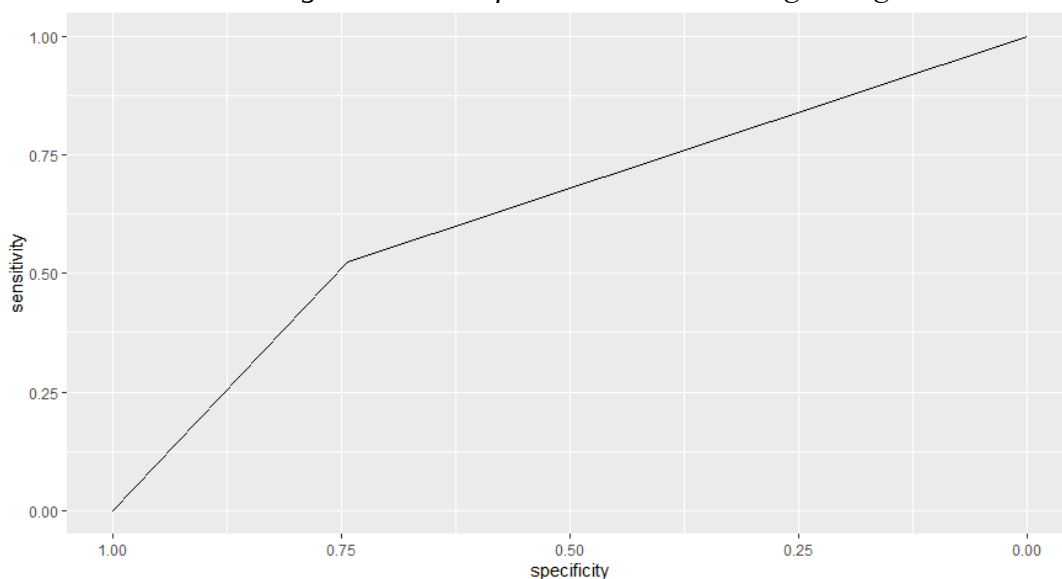
	Tényleges	
Előrejelzett	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	2 939	558
Nemteljesítés (1)	1 015	613

Forrás: saját szerkesztés

Az elemzés alapján a pontosság mutató (Accuracy) értéke 0,6931, ami azt jelenti, hogy 69,31%-ban helyesen osztályozta az ügyfeleket a döntési fa valószínűségi metszettel. Az érzékenység mutató (Sensitivity) 0,5235, ami azt mutatja, hogy a pozitív osztályba tartozó megfigyelések 52,35%-a helyesen lett azonosítva. A specificitás (Specificity) mutató értéke 0,7433, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések csak 74,33%-a lett helyesen azonosítva. A pozitív előrejelzési érték (Positive Predictive Value) 37,65%, míg a negatív előrejelzési érték (Negative Predictive Value) 84,04%.

Ebben az esetben a ROC görbét is a következő ábrán látható.

20. Ábra: ROC görbe: döntési fa előzetes valószínűségek megváltoztatásával



Forrás: saját szerkesztés

Az AUC vagyis a görbe alatti terület 0,6334, tehát a modell egy közepesen jó osztályozásba sorolható. Tehát az előzetes valószínűségek megváltoztatásával készült döntési fa esetén jobb eredményt kaptam, mint a veszteségmátrix bevonása esetén, de gyengébb eredményt, mint az alulmintavételezés során.

3.3.3. Véletlen erdő

A véletlen erdőt 500 döntési fa felhasználásával lett kialakítva és az OOB (out-of-bag) becslése a hibaaarányra 22,44%. Az out-of-bag becslés lényege, hogy a modellezés folyamán minden döntési fa csak azokat az adatokat használja fel a tanításhoz, amelyek nem kerültek bele az adott fa mintavételezésébe. Ezután az adott fa algoritmus kihagyott adatpontokon történő előrejelzéseit összesítik, és ezáltal kapnak egy átlagosan becslést a modell teljesítményére anélkül, hogy külön validációs adathalmazt kellene létrehozni (Dzik-Walczak et al., 2021). A 0 osztály esetében a klasszifikációs hiba 11,81% volt, míg az 1 osztály esetében ugyan ez a hiba 88,19% volt. A klasszifikációs mátrix a következő táblázatban látható.

12. Táblázat: Klasszifikációs mátrix: véletlen erdő

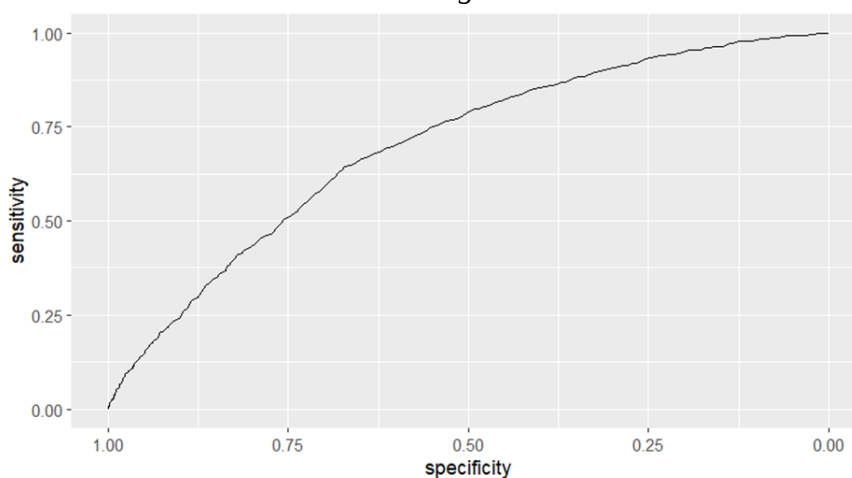
Előrejelzett	Tényleges	
	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	3 886	1 093
Nemteljesítés (1)	68	78

Forrás: saját szerkesztés

Az elemzés alapján a pontosság mutató (Accuracy) értéke 0,774, ami azt jelenti, hogy 77,40%-ban helyesen osztályozta az ügyfeleket a véletlen erdő. Az érzékenységi mutató (Sensitivity) 0,06661, ami azt mutatja, hogy a pozitív osztályba tartozó megfigyelések 6,66%-a helyesen lett azonosítva. A specifitási mutató értéke 0,9828, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések csak 98,28%-a lett helyesen azonosítva. A pozitív előrejelzési érték (Positive Predictive Value) 53,45%, míg a negatív előrejelzési érték (Negative Predictive Value) 78,04%.

Ezt követően meghatároztam a ROC görbét is, annak érdekében, hogy a modellek pontossága közötti verseny minél pontosabb és biztosabb eredményt mutasson. Az AUC vagyis a görbe alatti terület 0,7009, tehát a modell egy jó osztályozásba sorolható. A következő ábra a ROC görbét mutatja a véletlen erdő esetén.

21. Ábra: ROC görbe: véletlen erdő



Forrás: saját szerkesztés



2024. MÁJUS 15-18.

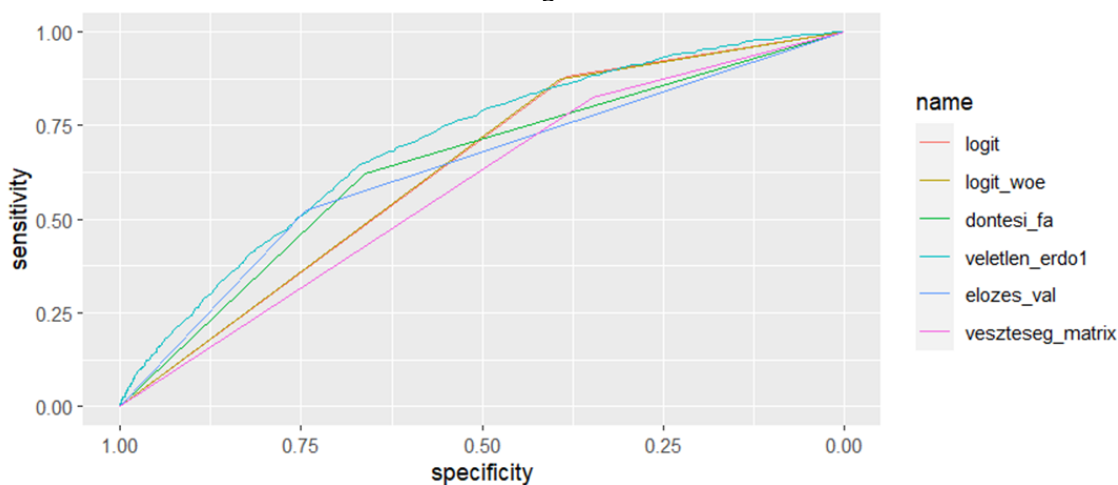
KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Annak érdekében, hogy jobban össze lehessen hasonlítani a modelleket az alábbi ábrán látható az összes általam becsült modell ROC görbéje.

22. Ábra: ROC görbék: összehasonlítás



Forrás: saját szerkesztés

A klasszifikációs mátrix, valamint a ROC görbe alapján arra az eredményre jutottam, hogy a véletlen erdő a legjobb modell a felsoroltak és az elemzettek közül.

13. Táblázat: Összefoglaló táblázat mutatókkal

Modell neve	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Logisztikus regresszió	49,68%	38,32%	88,04%	91,54%	29,71%	0,631
Logisztikus regresszió WoE transzformált	50,26%	39,25%	87,45%	91,35%	29,89%	0,633
Alulmintázott döntési fa	65,25%	62,25%	66,14%	35,25%	85,54%	0,642
Döntési fa a veszteségmátrix bevonásával	65,25%	62,25%	66,14%	35,25%	85,54%	0,586
Döntésifa előzetes valószínűségek megváltoztatásával	69,31%	52,35%	74,33%	37,65%	84,04%	0,633
Véletlen erdő	77,40%	6,66%	98,28%	53,45%	78,04%	0,700

Forrás: saját szerkesztés



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Következtetések

Egy Lending Club-ról származó adatbázis adatai alapján hasonlítottam össze a logisztikus regresszió teljesítményét, a döntési fával, valamint a véletlen erdővel. A vizsgált adatbázis 13.182 olyan ügyfelet tartalmazott, akik teljes mértékben visszafizették a hitelt és 3 904 olyat, akik nem teljesítették a hiteltörlesztést. Az adattisztítás után elvégzett elemzések arra irányultak, hogy különböző változók kapcsolatát vizsgáljam a nemteljesítés változóval. A változók közötti összefüggéseket t-próbával, valamint khi-teszttel vizsgáltam.

A tesztek alapján a következő változók befolyásolják a nemteljesítést: a hitel hossza (term), és a jelenlegi munkaviszony hossza (emp_length), a lakástulajdon (home_ownership), a jövedelem ellenőrzési állapota (verification_status), a minősítés (grade), az al-minősítés (sub_grade), a hitel összege (loan_amnt), hitelfelvevő összes adósságkötelezettségének aránya a havi törlesztőrészletben (dti), a FICO pontszám legkisebb értéke (fico_range_low), a FICO pontszám legnagyobb értéke (fico_range_high), a hitelfelvevő által felhasznált hitelösszeg az összes rendelkezésre álló ruvírozó hitelhez viszonyítva (revol_util), a hitelfelvevő összes rendelkezésre álló hitelkeretének felső határa (total_rev_hi_lim), a rendelkezésre álló, fel nem használt hitelkeret összegének aránya a bankkártyák esetében (bc_open_to_buy), a hitelfelvevő összes hitelszámla száma (total_cu_tl), a megkeresések száma az elmúlt 6 hónapban (inq_last_6mths), a hitelfelvevő által jelenleg használt aktív ruvírozó hitelszámlák száma (num_actv_rev_tl), valamint a hitelfelvevő által az elmúlt 12 hónapban megnyitott hitelszámlák száma (num_tl_op_past_12m).

Klasszifikációs mátrix és ROC görbe, valamint AUROC segítségével vizsgáltam a modellek teljesítményét. Ahogy az a tesztekben kiderült a véletlen erdő teljesített a legjobban, ez a modell tudta legnagyobb százalékban kiszűrni a nemteljesítőket, valamint legpontosabban előre jelezni a teljesítőket és nemteljesítőket.

A kutatás egyik korlátja a felhasznált adathalmaz minősége és mérete, amely befolyásolhatja az algoritmusok pontosságát és általánosíthatóságát. Az eredmények alátámasztják, hogy a különböző algoritmusok paraméterezése és finomhangolása szintén



jelentős hatással van az eredményekre, amelyeket a hitelminősítési folyamatok optimalizálására és automatizálására lehet felhasználni, különösen a pénzügyi intézményeknél, amelyek számára fontos a kockázatkezelés és a hitelképesség pontos értékelése. A mesterséges intelligencia alapú algoritmusok alkalmazása növelheti a prediktív pontosságot és hatékonyságot, ezáltal javítva a pénzügyi döntéshozatal minőségét.



Irodalomjegyzék

Aleksandrova, Y. (2021). Comparing performance of machine learning algorithms for default risk prediction in peer to peer lending. *TEM Journal*, 10(1), 133-143, DOI:[10.18421/TEM101-16](https://doi.org/10.18421/TEM101-16)

Ahmed, S. E. (2023). Determinants of Credit Ratings and Comparison of the Rating Prediction Performances of Machine Learning Algorithms. In *E3S Web of Conferences*, 409, 05013. EDP Sciences. DOI:10.1051/e3sconf/202340905013

Ágoston, N. (2022). Külföldi csődelőrejelző módszerek szisztematikus irodomelemzése. *Vezetéstudomány*, 53(1), 69-89. <https://doi.org/10.14267/VEZTUD.2022.01.06>

Berger, E. A., Butler, A. W., Hu, E., & Zekhnini, M. (2021). Financial integration and credit democratization: Linking banking deregulation to economic growth. *Journal of Financial Intermediation*, 45, 100857. <https://doi.org/10.1016/j.jfi.2020.100857>

Brimacombe, M. (2016). Large sample convergence diagnostics for likelihood based inference: Logistic regression. *Statistical Methodology*, 33, 114-130., DOI:10.1016/j.stamet.2016.08.001

Crook, J., & Banasik, J. (2012). Forecasting and explaining aggregate consumer credit delinquency behavior. *International Journal of Forecasting*, 28(1), 145-160. <https://doi.org/10.1016/j.ijforecast.2010.12.002>

Chang, A. H., Yang, L. K., Tsaih, R. H., & Lin, S. K. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data. *Quantitative Finance and Economics*, 6(2), 303-325, DOI:[10.3934/QFE.2022013](https://doi.org/10.3934/QFE.2022013)

Davaadorj, Z., Enkhtaivan, B., & Lu, W. (2024). The role of job titles in online peer-to-peer lending: An empirical investigation on skilled borrowers. *Journal of Behavioral and Experimental Finance*, 41, 100890., DOI:[10.1016/j.jbef.2024.100890](https://doi.org/10.1016/j.jbef.2024.100890)

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178-1192., DOI:[10.1016/j.ejor.2021.06.053](https://doi.org/10.1016/j.ejor.2021.06.053)

Dzik-Walczak, A., & Heba, M. (2021). An implementation of ensemble methods, logistic regression, and neural network for default prediction in Peer-to-Peer lending. *Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu*, 39(1), 163-197, DOI:[10.18045/zbefri.2021.1.163](https://doi.org/10.18045/zbefri.2021.1.163)



2024. MÁJUS 15-18.

KOLOZSVÁR

XXVII. ERDÉLYI TUDOMÁNYOS DIÁKKÖRI KONFERENCIA REÁL- ÉS HUMÁNTUDOMÁNYOK

CONFERINȚA ȘTIINȚIFICĂ STUDENȚEASCĂ DIN TRANSILVANIA, EDIȚIA A 27-A
27TH TRANSILVANIA STUDENTS' SCIENTIFIC CONFERENCE
ȘTIINȚE REALE ȘI UMANISTE // FORMAL AND EMPIRICAL SCIENCES

Estran, R., Souchaud, A., & Abitbol, D. (2022). Using a genetic algorithm to optimize an expert credit rating model. *Expert Systems with Applications*, 203, 117506., DOI:10.1016/j.eswa.2022.117506

Lyócsa, Š., Vašaničová, P., Hadji Misheva, B., & Vateha, M. D. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation*, 8(1), 32, DOI:[10.1186/s40854-022-00338-5](https://doi.org/10.1186/s40854-022-00338-5)

Norbert, Á. (2022). Mesterséges intelligencia és gépi tanulási módszerek a vállalati fizetéseképtelenség becslésére. *Statisztikai Szemle*, 100(6), 584-609.

Nyitrai, T., & Virág, M. (2017). Magyar vállalkozások felszámolásának előrejelzése pénzügyi mutatóik idősorai alapján. *Közgazdasági Szemle*, 64(3), 305-324., DOI:[10.18414/KSZ.2017.3.305](https://doi.org/10.18414/KSZ.2017.3.305)

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621-4631, DOI:[10.1016/j.eswa.2015.02.001](https://doi.org/10.1016/j.eswa.2015.02.001)

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, 113986, DOI:[10.1016/j.eswa.2020.113986](https://doi.org/10.1016/j.eswa.2020.113986)

Fraisse, H., & Laporte, M. (2022). Return on investment on artificial intelligence: The case of bank capital requirement. *Journal of Banking & Finance*, 138, 106401., DOI:[10.1016/j.jbankfin.2022.106401](https://doi.org/10.1016/j.jbankfin.2022.106401)

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009-1029, DOI:[10.1111/fima.12295](https://doi.org/10.1111/fima.12295)

Kristóf, T., & Virág, M. (2020). A comprehensive review of corporate bankruptcy prediction in Hungary. *Journal of Risk and Financial Management*, 13(2), 35., DOI:[10.3390/jrfm13020035](https://doi.org/10.3390/jrfm13020035)

Kim, A., & Cho, S. B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering applications of Artificial intelligence*, 81, 193-199, DOI:[10.1016/j.engappai.2019.02.014](https://doi.org/10.1016/j.engappai.2019.02.014)

Kristóf, T., & Virág, M. (2020). A comprehensive review of corporate bankruptcy prediction in Hungary. *Journal of Risk and Financial Management*, 13(2), 35, DOI:[10.3390/jrfm13020035](https://doi.org/10.3390/jrfm13020035)



Pang, S., Hou, X., & Xia, L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change*, 165, 120462, DOI:[10.1016/j.techfore.2020.120462](https://doi.org/10.1016/j.techfore.2020.120462)

Prisznyák, A. (2022). Bankrobotika: mesterséges intelligencia és gépi tanulás alapú banki kockázatkezelés: Pénzmosás és terrorizmusfinanszírozás megakadályozása. *Pénzügyi szemle*, 67(2). https://doi.org/10.35551/PFQ_2022_2_8

Teply, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance*, 51, 100904, DOI:[10.1016/j.najef.2019.01.001](https://doi.org/10.1016/j.najef.2019.01.001)

Uzonyi, N. (2021). Gépi tanulás és bestsellerek. *Mesterséges Intelligencia*, 3(2), 43-53, DOI:[10.35406/MI.2022.2.9](https://doi.org/10.35406/MI.2022.2.9)

Song, Y., Wang, Y., Ye, X., Zaretzki, R., & Liu, C. (2023). Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *Information Sciences*, 629, 599-617. <https://doi.org/10.1016/j.ins.2023.02.014>

Shi, B., Zhao, X., Wu, B., & Dong, Y. (2019). Credit rating and microfinance lending decisions based on loss given default (LGD). *Finance Research Letters*, 30, 124-129, DOI:[10.1016/j.frl.2019.03.033](https://doi.org/10.1016/j.frl.2019.03.033)

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy*, 24(6), 801, DOI:[10.1016/j.physa.2019.122370](https://doi.org/10.1016/j.physa.2019.122370)

Kgoroadira, R., Burke, A., Di Pietro, F., & van Stel, A. (2023). Determinants of firms' default on unsecured loans in the P2P crowdfunding market. *Journal of International Financial Markets, Institutions and Money*, 89, 101882., DOI:[10.1016/j.intfin.2023.101882](https://doi.org/10.1016/j.intfin.2023.101882)

Wang, D., Chen, Z., Florescu, I., & Wen, B. (2023). A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance*, 64, 101869., DOI: [10.1016/j.ribaf.2022.101869](https://doi.org/10.1016/j.ribaf.2022.101869)

Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia computer science*, 199, 1128-1135, DOI:[10.1016/j.procs.2022.01.143](https://doi.org/10.1016/j.procs.2022.01.143)

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1), 223-230, DOI:[10.1016/j.eswa.2010.06.048](https://doi.org/10.1016/j.eswa.2010.06.048)



Wu, Y., & Pan, Y. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. *Complexity*, 2021, 1-12, DOI:[10.31289/perspektif.v11i3.7180](https://doi.org/10.31289/perspektif.v11i3.7180)

Woo, H., & Sohn, S. Y. (2022). A credit scoring model based on the Myers–Briggs type indicator in online peer-to-peer lending. *Financial Innovation*, 8(1), 42, DOI:[10.1186/s40854-022-00368-z](https://doi.org/10.1186/s40854-022-00368-z)

Ye, X., Dong, L. A., & Ma, D. (2018). Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, 32, 23-36, DOI:[10.1016/j.procs.2019.12.017](https://doi.org/10.1016/j.procs.2019.12.017)

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. *Scientific Reports*, 11(1), 18759, DOI:[10.1038/s41598-021-98361-6](https://doi.org/10.1038/s41598-021-98361-6)

Zhang, M., Guo, X., Lu, X., & Jiang, Y. (2024). Credit Guarantee, Financing Structure, and Firm ESG Performance. *Finance Research Letters*, 105315., DOI:[10.1016/j.frl.2024.105315](https://doi.org/10.1016/j.frl.2024.105315)

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534, 122370, DOI:[10.1016/j.physa.2019.122370](https://doi.org/10.1016/j.physa.2019.122370)

Zhou, Y., Shen, L., & Ballester, L. (2023). A two-stage credit scoring model based on random forest: Evidence from Chinese small firms. *International Review of Financial Analysis*, 89, 102755., DOI:[10.1016/j.irfa.2023.102755](https://doi.org/10.1016/j.irfa.2023.102755)