Közgazdász Fórum
Forum on Economics and Business
26(149), 58–86.

2023/1

Comparison of credit scoring algorithms: artificial intelligence-based algorithms vs. traditional methods

ZÉTÉNY SZILÁGYI-NAGY

The aim of the study is to compare the performance of logistic regression, decision tree and random forest on a real P2P (peer-to-peer) lending database. The peer-to-peer (or person-to-person, P2P) lending market facilitates financial transactions between borrowers and lenders. Credit risk management is traditionally based on credit scoring, which is used to estimate the probability of default on a loan. An appropriate credit-scoring technique is very important for the long-term success of both financial institutions and P2P lending platforms. The comparison of credit scoring algorithms is based on Lending Club data. The database covers the first quarter of 2017 and contains 563 847 observations and 23 different variables. The results obtained suggest that random forest is the better credit scoring algorithm for the Lending Club database.

Keywords: credit score, peer-to-peer lending, logistic regression, random forest, decision tree.

JEL codes: G21, G23, D14.

Kiadó: Romániai Magyar Közgazdász Társaság, a Babeş–Bolyai Tudományegyetem Közgazdaság- és Gazdálkodástudományi Karának Magyar Intézete és a Partiumi Keresztény Egyetem
ISSN: 1582-1986 www.kozgazdaszforum.ro

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

SZILÁGYI-NAGY ZÉTÉNY[1]

A tanulmány célja összehasonlítani a logisztikus regresszió, döntési fa és a véletlen erdő teljesítményét egy valós P2P (peer-to-peer) hitelezési adatbázison. A peer-to-peer (vagy személyközi, P2P) hitelezési piac megkönnyíti a hitelfelvevők és a hitelezők közötti pénzügyi tranzakciókat. A hitelkockázat-kezelés hagyományosan a hitelpontozáson alapszik, amely segítségével a hitelek nemteljesítésének valószínűségét becsüljük. A megfelelő hitelminősítési technika nagyon fontos mind a pénzügyi intézmények, mind a P2P hitelezési platformok hosszú távú sikere szempontjából. A hitelminősítő algoritmusok összehasonlítása a Lending Clubról származó adatok alapján történik. Az adatbázis 2017 első negyedévére terjed ki, 563 847 megfigyelést és 23 különböző változót tartalmaz. Az eredmények alapján arra a következtetésre jutunk, hogy a véletlen erdő a jobb hitelminősítő algoritmus a Lending Club adatbázisa esetében.

Kulcsszavak: hitelminősítés, személyközi hitelezés, logisztikus regresszió, véletlen erdő, döntési fa.

JEL kódok: G21, G23, D14.

Bevezető

Teply és Polena (2020) szerint a FinTech pénzügyi innovációjaként a személyközi (peer-to-peer, P2P) hitelezés egy új, online pénzügyi közvetítés, amely a hitelfelvevőket köti össze a 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 az online jelenlétük miatt a működési költségeik alacsonyabbak, mint a fizikai bankoké. Ez teszi lehetővé a versenyképesebb feltételek kínálását a hitelfelvevők és a hitelezők számára. A hitelfelvevők átlagosan alacsonyabb kamatlábakat fizetnek a P2P-hitelezési platformokon, mint a bank által nyújtott hitelek esetén. A jól diverzifikált hitelportfólióval rendelkező hitelezők magasabb hozamot érnek el, mint a hagyományos megtakarítási számlákon. Ezek a tények egyre inkább népszerűvé teszik a P2P-hitelezést mind a hitelfelvevők, mind a hitelezők számára.

[1] BSc-hallgató, Babeș–Bolyai Tudományegyetem, Közgazdaság- és Gazdálkodástudományi Kar, e-mail: zeteny.szilagyi@stud.ubbcluj.ro.

Fontos kiemelni a P2P-hitelezés korlátait is ahhoz, hogy a hitelezési piacról egy átfogó képet kapjunk. Zhao et al. (2021) szerint az egyik jelentős korlát a kockázatkezeléssel kapcsolatos, hiszen a P2P-hitelezési platformokon a hitelfelvevők hitelképességének értékelése gyakran kevésbé alapos, mint a hagyományos bankok esetében, ami magasabb kockázatot jelent a hitelezők számára. A hagyományos bankokkal ellentétben a hitelkockázatot nem a P2P-platformnak, hanem a befektetőknek kell viselniük (Stern et al. 2017). A P2P-hitelezés likviditási kockázata is fontos, a befektetők nem kapják azonnal vissza a befektetett összeget, meg kell várniuk a hitel futamidejének a végét. A kamatkockázat és a platformkockázat is korlátja a P2P-hitelezésnek, hiszen a P2P-hitelezési platformok üzleti modelljei és pénzügyi helyzete változhat, ami a befektetők számára kockázatot jelent. Ezek mellett a platform meghibásodásából, csalásból vagy kiberbűnözésből eredő potenciális kockázati tényezők a befektetőkre és a platformokra is kockázatot jelentenek (Milne–Parboteeah 2016).

A legelső P2P-hitelezési platformot, a Zopát 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 legfontosabb P2P-hitelezési platform, amelyet 2014-ben bevezettek a New York-i tőzsdére. A P2P-hitelezés piacának értéke 2020-ban 84,89 milliárd dollár volt, és az előrejelzések szerint 2028-ra eléri az 578,03 milliárd dollár értéket.

Az egyedi hitelkérelmek növekedésével a hitelkockázat értékelése egyre fontosabbá vált a pénzügyi szakemberek és kutatók számára is. Ágoston (2022a) szerint a hagyományos módszerek, amelyek sokáig dominálták ezt a területet, egyre nagyobb 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 tanulmányban áttekintem az új megközelítéseket, majd hatékonyság és alkalmazhatóság szempontjából összehasonlítom ezeket a hagyományos módszerekkel. Az elemzés célja, hogy betekintést nyújtson abba, hogy miként lehetnek az új technológiák hatékony eszközök a hitelminősítő intézmények számára a kockázatok kezelésében és a döntéshozatalban. A klasszifikációs mátrix alapján felépített mérőszámok segítségével összehasonlítom a döntési fa, a véletlen erdő, valamint a logisztikus regresszió teljesítményét. A modellek teljesítményének értékelése alapján arra a kérdésre kapunk választ, hogy az új technológiák képe-

sek-e javítani a hitelminősítő intézmények hatékonyságát és eredményességét a hitelezési döntéshozatal folyamatában.

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

Szakirodalmi áttekintés

A hagyományos banki hitelezéssel szemben a P2P-hitelezés újabb megközelítést jelent, amely során közvetlen kapcsolat alakul ki a kölcsönt igénylők 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. A döntéshozóknak fontos megérteniük mindkét lehetőség előnyeit és korlátait annak érdekében, hogy hatékonyan tudjanak választani a finanszírozási lehetőségek között. A hitelezési döntés hasonló a két finanszírozási forrás esetén (Kgoroeadira 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. Pang et al. (2021) szerint a hitelminősítő pontszámok lehetővé teszik a hitelezők számára, hogy meghatározzák azt a valószínűséget, amellyel a hitelfelvevő vissza fogja fizetni a kölcsönt a megállapított feltételek szerint. A hitelezők és a hitelfelvevők számára egyaránt fontos, hogy ezek a hitelminősítések megbízhatóak és pontosak legyenek.

Számos tanulmány vizsgálja a hitelminősítő algoritmusokat és hasonlítja össze teljesítményüket (például Fraisse–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 módszereket és a mesterséges intelligencia alapú hitelminősítő algoritmusokat. 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 eredményei könnyebben értelmezhetőek, valamint magyarázhatóak, mivel matematikai modellek alapján működnek.

A hagyományos hitelminősítő algoritmusok az 1970-es években kezdtek elterjedni. Ezek az algoritmusok általában statisztikai modelleken és számítógépes programokon alapulnak, amelyek az ügyfelek pénzügyi adatait és egyéb tényezőket elemzik a hitelképességük meghatározása során. 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 hitelkockázat területén leggyakrabban használt statisztikai módszerek a logisztikus regresszió és a Lasso regresszió. Dumitrescu et al. (2022) szerint a logisztikus regresszió valószínűségi alapú osztályozást alkalmaz arra, 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 pontosságának javítása a cél, valamint annak megállapítása, hogy a hitel késedelmes-e.

A mesterséges intelligencia alapú hitelminősítő algoritmusok az elmúlt 10 évben terjedtek el (Zhou et al. 2019). Ezek az algoritmusok összetettebb adatelemzést végeznek, és képesek nagyobb adatmennyiségeket és változatosabb adatforrásokat feldolgozni, mint például a közösségi média aktivitás vagy az online vásárlási szokások. 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.

Dzik-Walczak és Heba (2021) összehasonlította a logisztikus regresszió és a döntési fa teljesítményét. Az elemzésben használt adatbázis a Lending Club oldaláról származik, a 2011 és 2013 közötti periódusra vonatkozik. Megállapították, hogy a logisztikus regresszió jobb eredményt ért el, mint a 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.

Wang et al. (2022) a magánszemélyek hitelkockázati jellemzőit vizsgálták logisztikus regresszióval, valamint XGBooston alapuló diszkriminációs modellel. A nemteljesítések közötti megkülönböztetések, a minél pontosabb eredmény eléréséhez, XGBoosttal történtek. A logisztikus regresszió esetén volt a legmagasabb az AUC értéke és a legnagyobb pontossággal rendelkezett. Arra a következtetésre jutottak, hogy az XGBooston alapuló hitelkockázat-értékelési modell nagyon jó nemteljesítési megkülönböztető képességgel és robusztussággal rendelkezik.

Zhou et al. (2019) egy véletlen erdő algoritmuson alapuló osztályozási módszert vizsgáltak, majd hasonlították össze az eredményeit egy XGBoost algoritmussal. Ha a hitelező nagyon alacsony elfogadási szintet alkalmaz, vagyis nagyon kis kockázatot vállal, akkor a véletlen erdő hatékonysága csökken. Ez azt jelenti, hogy a módszer kevésbé lesz pontos vagy megbízható az alacsonyabb elfogadási szinteken. A költségérzékenység figyelembevétele esetén sem sikerült jelentősen javítani a módszer nemteljesítő hitelfelvevők azonosítására vonatkozó képességét. Ágoston (2022b) az SVM-módszert hasonlította össze a neurális hálóval, valamint a véletlen erdővel. Az AdaBoostot és a Bagginget használta az adatok kiegyensúlyozásához, elemzéséhez, valamint rendezéséhez. Az SVM több mint 5%-kal magasabb osztályozási pontosságot ért el az elemzett adatbázison. Megállapította, hogy a csődelőrejelzésben a fejlődés leginkább a mesterséges intelligenciának, valamint a gépi tanulásnak köszönhető.

A hitelminősítés területén egyre nagyobb szerepet kapnak a modern algoritmusok, mint például a neurális hálózatok, transzformerek és természetes nyelvfeldolgozási (NLP) modellek, mivel jelentős előrelépést jelentenek a hagyományos statisztikai és gépi tanulási módszerekhez képest. A neurális hálózatok képesek komplex mintázatok felismerésére, ami hasznos a hitelminősítés során, ahol számos változó kölcsönhatását kell figyelembe venni. A mélytanulási technikák, mint a visszacsatolt neurális hálózatok, lehetővé teszik nagy mennyiségű adat feldolgozását és az adatok közötti rejtett összefüggések feltárását, ami különösen előnyös a nemlineáris kapcsolatok felismerésében.

Wang és Xiao (2022) szerint a transzformerek, amelyeket eredetileg a természetes nyelvfeldolgozásban alkalmaztak, egyre inkább beépülnek a hitelminősítés területére is. Ezek a modellek hatékonyan kezelik a szekvenciális adatokat és képesek hosszú távú kapcsolatok azonosítására, ami lehetőséget biztosít arra, hogy a hitelminősítési folyamatban figyelembe vegyék az időben változó adatokat, például az ügyfél hiteltörténetének alakulását.

Az NLP-modellek különösen hatékonyak, amikor nem strukturált adatokat, például szöveges értékeléseket vagy ügyfélpanaszokat kell elemezni. Ezek a modellek lehetővé teszik, hogy a hitelminősítési folyamatba integrálják az ilyen típusú információkat is, ezáltal növelve a modellek előrejelző képességét. Az NLP-technikák segítségével a szöveges adatok elemzése révén még teljesebb képet kaphatunk egy hitelfelvevő kockázatosságáról Alonso Robisco és Carbo Martinez (2022) alapján.

Az elmúlt években a kutatások arra a következtetésre jutottak, hogy számos hitelkockázat-értékelési területen a mélytanulás felülmúlja a hagyományos gépi tanulási módszereket, és az osztályozók együttesen jelentősen jobban teljesítenek, mint az egyes osztályozók (Shen et al. 2021). Xia et al. (2017) kutatása az együttes gépi tanulásra épül, amely több technika kombinációja, amelynek teljesítménye jobb, mint az egyes technikáké külön-külön. Ezen túl arra is kitérnek, hogy ezek a modellek hogyan használhatók fel hatékonyan a hitelminősítésben. Emellett a hibrid modellek, amelyek többféle algoritmust ötvöznek, pontosságuk miatt egyre népszerűbbek.

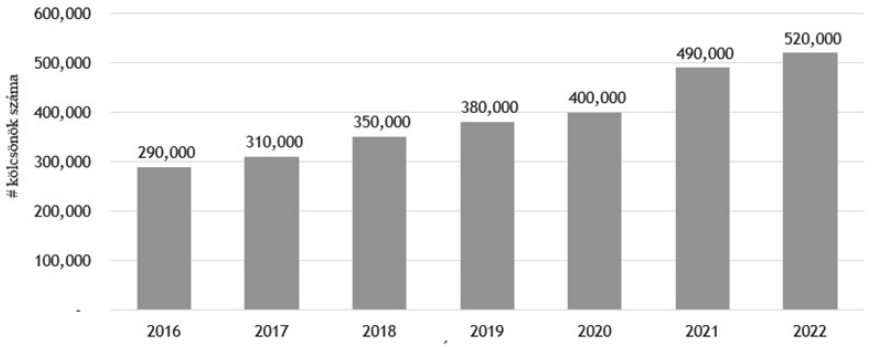
Rudin (2019) a modern hitelminősítő modellek átláthatóságának és magyarázhatóságának a fontosságát elemzi, főként a mélytanulási algoritmusok esetében, amelyek gyakran fekete doboz módszereket alkalmaznak. Kiemeli a magyarázható mesterséges intelligencia (XAI) jelentőségét, amely növelheti a hitelminősítési modellekbe vetett bizalmat.

Adatok és módszertan

Ebben a fejezetben bemutatom az elemzésben használt adatbázist és az adattisztítási folyamatot, amely magába foglalja a nem releváns változók törlését, a változók transzformációját, a hiányzó és kiugró értékek kezelését és az alulmintavételezést. Az adatok bemutatása után az elemzésben használt módszerekre is kitérek.

Wang et al. (2021) szerint a Lending Club egy olyan peer-to-peer hitelnyújtó vállalat, amely egy online platformon keresztül hozza össze a hitelfelvevőket a befektetőkkel. Olyan személyeknek nyújt hitelt, akiknek 1000 és 40 000 dollár közötti személyi kölcsönre van szükségük. A hitelfelvevők a kölcsön teljes összege és a kezelési költség közötti különbséget megkapják. 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 1. ábrán látható, hogy az elmúlt években hány új kölcsön került fel a Lending Club oldalára. Az adatokból látszik a P2P-hitelezés dinamikus fejlődése.



Forrás: Saját szerkesztés a Lending Club adatai alapján

1. ábra: A Lending Club által folyósított hitelek száma, 2016–2022

Az elemzés a Lending Club 2017-es első negyedévének hiteladataira épül. Az adatbázis 58 102 hitelfelvevő adatait tartalmazta. Az eredeti adatkészlet 117 változót tartalmazott, amelyből 35, az elemzés szempontjából releváns változót választottam ki. Az 1. táblázatban láthatók a változók nevei, leírásuk, hogy át lettek-e alakítva vagy sem, valamint hogy tartalmaztak-e kiugró vagy hiányzó értékeket.

1. táblázat: A változók leírása

Átalakítva	Változó neve	Leírás	Mérték-egység
Nem	loan_amnt (hitel összege)	A hitel összege.	USD
Igen	term (futamidő)	Hitelezési periódus hossza: 36 vagy 60 hónap.	Hónap
Igen	grade (minősítés)	Minősítési kategória: A, B, ..., F.	–
Igen	sub_grade (al-minősítés)	Minősítési alkategória: az „A”-tól „G”-ig terjedő minősítések további alosztályokra bontása, például „A1”, „A2” stb.	–
Igen	emp_length (régiség)	Munkaviszony hossza a jelenlegi munkahelyen.	Év

Átalakítva	Változó neve	Leírás	Mérték-egység
Igen	home_ownership (lakástulajdon)	A lakástulajdoni státusz: bérelt, saját, jelzálog, egyéb.	–
Igen	annual_income (jövedelem)	A hitelfelvevő éves jövedelme.	USD
Igen	verification_status (ellenőrzési állapot)	A hitelfelvevő jövedelmének ellenőrzési állapota: ellenőrzött, részben ellenőrzött, nem ellenőrzött.	–
Igen	loan_status (nemteljesítés)	A hitel jelenlegi állapota: folyamatban, késedelmes, elfogadva, visszautasítva.	–
Igen	purpose (cél)	A hitel célja.	–
Igen	addr_state (állam)	A hitelfelvevő lakhelyének állama.	–
Nem	delinq_2yrs (30 nap késedelmek)	A 30 napon túli késedelmes esetek száma az elmúlt 2 évben.	Darab
Igen	dti (adósságjövedelem arány)	A hitelfelvevő havi jövedelméhez viszonyított teljes adósságának aránya.	%
Nem	inq_last_6mths (megkeresések száma 6 hónapra)	Megkeresések száma az elmúlt 6 hónapban.	Darab
Igen	earliest_cr_line (legkorábbi hitelkeret nyitás)	A hitelfelvevő legkorábbi bejelentett hitelkerete megnyitásának napja.	Dátum
Igen	open_acc (számlák száma)	A számlák száma.	Darab
Nem	pub_rec (negatív események száma)	Nyilvános nyilvántartásokban szereplő negatív események száma.	Darab
Nem	acc_now_delinq (késedelmes számlák)	A jelenleg késedelmes számlák száma.	Darab
Igen	total_rev_hi_lim (hitelkeret felső határa)	A hitelfelvevő összes rendelkezésre álló hitelkeretének felső határa.	USD
Igen	total_cu_tl (hitelszámla számok)	A hitelfelvevő összes hitelszámlájának száma.	Darab
Nem	bc_open_to_buy (hitelkeret aránya)	A rendelkezésre álló, fel nem használt hitelkeret összegének aránya a bankkártyák esetében.	USD
Igen	mo_sin_rcnt_rev_tl_op (rulírozó hitelkeret megnyitás napja)	A legutóbbi rulírozó hitelszámla megnyitása óta eltelt hónapok száma.	Hónap
Nem	num_actv_rev_tl (aktív rulírozó hitelek)	A hitelfelvevő által jelenleg használt aktív rulírozó hitelszámlák száma.	Darab

Átalakítva	Változó neve	Leírás	Mérték-egység
Nem	num_il_tl (nem rulírozó hitelszámlák)	A hitelfelvevő által használt egyéb hitelszámlák (nem rulírozó) száma.	Darab
Nem	num_op_rev_tl megnyi-tott rulírozó hitelek)	A hitelfelvevő által megnyitott rulírozó hitelszámlák száma.	Darab
Igen	revol_util (rulírozó hitelhez felhasznált hitelösszeg-arány)	A hitelfelvevő által felhasznált hitelösszeg az összes rendelkezésre álló rulírozó hitelhez viszonyítva.	%
Nem	num_tl_op_past_12m (egy éven belül megnyitott hitelszámlák)	A hitelfelvevő által az elmúlt 12 hónapban megnyitott hitelszámlák száma.	Darab
Nem	percent_bc_gt_75 (hitelkeret-kihasználtság 75% felett)	A bankkártyák aránya, amelyek esetében a hitelkeret kihasználtsága meghaladja a 75%-ot.	%
Nem	pub_rec_bankruptcies (csődök száma)	A nyilvános nyilvántartásokban szereplő csődök száma.	Darab
Nem	total_acc (hitelkeret száma)	A hitelfelvevő hitelkereteinek száma.	Darab
Nem	chargeoff_within_12_ months (12 hónapban lévő hitelkiesések)	A 12 hónapon belüli hitelkiesések száma.	Darab
Igen	delinq_amnt (késedelmi számlák összege)	Azoknak a számláknak a lejárt összege, amelyek kifizetésével a hitelfelvevő késik.	USD
Nem	tax_liens (adótartozások)	Adótartozások száma, amely jelzi, hogy hányszor jegyeztek be adótartozásokat, 14 különböző típusa van.	Darab
Nem	fico_range_low (legkisebb FICO-érték)	A FICO-pontszám* legkisebb értéke.	Pontszám
Nem	fico_range_high (legnagyobb FICO-érték)	A FICO-pontszám legnagyobb értéke.	Pontszám

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

Forrás: A Kaggle-ről származó változók magyarázata

Minden változó esetén megvizsgáltam a kiugró értékeket, valamint a hiányzó értékeket. Egyik változó sem tartalmazott hiányzó értékeket, viszont kiugró értékeket találtam több változó esetében is. A kiugró értékek kezelése fontos lépés az adatelemzés során, mivel ezek jelentősen befolyásolhatják a modellezés eredményét. A kiugró értékek azok az adatok, amelyek jelentősen eltérnek a többi adatponttól, és gyakran torzíthatják a statisztikai elemzéseket, például a modellek pontosságát

és megbízhatóságát. Kiugró értékeknek azokat az értékeket tekintettem, amelyek meghaladják a felső kvartilis (Q3) és az interkvartilis terjedelem (IQR) másfélszeresének az összegét. A jövedelem változó esetében a 170 900 USD feletti értékeket szűrtem ki (2913 megfigyelés). Az adózáskötelezettségi arány a havi törlesztőrészlet változó esetében a 44,39% feletti értékeket, a számlák száma változó esetében a 23 darab feletti értékeket szűrtem ki. A rulírozó hiteleknél felhasznált hitelösszeg esetében a 122,05 USD volt a küszöbérték, a hitelező hitelkereteinek felső határa esetében pedig a 82 800 USD. A hitelfelvevő hitelszámlának száma esetében az 5 darab hitelszámla feletti értékeket szűrtem ki. A 2. táblázat tartalmazza a változókat, a küszöbértéket a kiugró értékek esetén, és az ezt meghaladó megfigyelések számát. A kiugró értékeket kitöröltem az adatbázisból.

2. táblázat: Kiugró értékek elemzése

Változó	Küszöbérték	Mértékegység	Megfigyelések száma
annual_inc (jövedelem)	170 900	USD	2913
dti (adósság arány havi törlesztéssel)	44,391	%	758
open_acc (számlák száma)	23	Darab	2213
revol_util (rulírozó hitelhez felhasznált hitelösszegarány)	122,054	USD	3
total_rev_hi_lim (hitelkeret felső határa)	82 800	USD	2579
total_cu_tl (hitelszámlaszámok)	5	Darab	3310
delinq_amnt (késedelmi számlák száma)	10	Darab	143
mo_sin_rcnt_rev_tl_op (rulírozó hitelkeret megnyitásának napja)	39	Hónap	3149

Forrás: Saját szerkesztés

A mennyiségi változók leíró statisztikáit a 3. táblázat tartalmazza.

A hitel státusza változónak hat értéke volt az eredeti adatbázisban (teljesen visszafizetve, leírásra került, folyamatban levő, türelmi időben levő, 16–30 nap közötti késedelemmel rendelkezik, 31–120 nap közötti késedelemmel rendelkezik), az elemzés szempontjából ezek közül kettő volt releváns (a teljesen visszafizetett és a nemteljesítő hitelek), a másik négy értéket eltávolítottam az adatbázisból. A hitel státusza változóból létrehoztam így a nemteljesítés változót. A nemteljesítés változó bináris változó: a 0 érték a teljesítő hiteleket, az 1-es érték a nemteljesítő hiteleket jelöli, amint az a 4. táblázatban is látható.

Az adattisztítás után az adatbázis 13 182 olyan ügyfél adatát tartalmazta, aki visszafizette a hitelt, illetve 3904 nemteljesítő ügyfél adatát.

3. táblázat: A mennyiségi változók leíró statisztikái

Változó	Min	Medián	Átlag	Max	Mérték-egység
loan_amnt	1000,000	12 000,000	14 180,000	40 000,000	USD
annual_income	5000,000	64 000,000	69 433,000	170 000,000	USD
loan_status	0,000	0,000	0,231	1,000	Kategória
delinq_2yrs	0,000	0,000	0,322	25,000	Darab
dti	0,000	17,422	17,983	44,333	%
inq_last_6mths	0,000	0,000	0,491	5,000	Darab
open_acc	1,000	10,000	10,000	23,000	Darab
pub_rec	0,000	0,000	0,254	47,000	Darab
acc_now_delinq	0,000	0,000	0,010	1,000	Darab
total_rev_hi_lim	300,000	23 300,000	27 342,000	82 800,000	USD
total_cu_tl	0,000	0,000	0,941	5,000	Darab
bc_open_to_buy	0,000	5650,000	9644,000	78 827,000	USD
mo_sin_rcnt_rev_tl_op	0,000	7,000	10,051	39,000	Hónap
num_actv_rev_tl	0,000	5,000	5,172	21,000	Darab
num_il_tl	0,000	6,000	8,123	81,000	Darab
num_op_rev_tl	1,000	7,000	7,732	23,000	Darab
revol_util	0,000	46,300	46,894	116,200	%
num_tl_op_past_12m	0,000	2,000	2,321	20,000	Darab
percent_bc_gt_75	0,000	33,300	38,832	100,000	%
pub_rec_bankruptcies	0,000	0,000	0,162	5,000	Darab
total_acc	2,000	21,000	22,611	87,000	Darab
chargeoff_within_12_months	0,000	0,000	0,070	6,000	Darab
delinq_amnt	0,000	0,000	0,000	0,000	USD
tax_liens	0,000	0,000	0,060	46,000	Darab
fico_range_low	660,000	690,000	696,100	845,000	Pontszám
fico_range_high	664,000	694,000	700,100	850,000	Pontszám

Forrás: Saját számítás

4. táblázat: Hitelek száma az átalakítások előtt és után

Változó	Hitelek száma
Hitel státusza	
Teljesen visszafizetve	13 182
Leírásra került	3 904
Folyamatban levő	24 952
Türelmi időben lévő	238
16–30 nap közötti késedelemmel rendelkezik	115
31–120 nap közötti késedelemmel rendelkezik	643
Összesen	**43 034**
Nemteljesítés	
Teljesítő (0)	13 182
Nemteljesítő (1)	3 904
Összesen	**17 086**

Forrás: A Lending Club oldaláról származó adatok alapján saját szerkesztés

Módszertan

Az adatbázis tisztítása után a logisztikus regresszió, a döntési fa, valamint a véletlen erdő teljesítményét hasonlítottam össze.

Logisztikus regresszió

Chang et al. (2022) szerint a leggyakrabban használt 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 bináris logisztikus regresszió alakja a következő:

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

ahol y egy bináris függő változó, amely értéke 0 vagy 1. Az x-ek a független változók, amelyeket a T függvény leképez valós számokra 0 és 1 között, úgy, hogy a becsült érték [0,1] közötti legyen. Az ε a hibatagot jelöli. A következő egyenlet mutatja, hogy a T függvény hogyan képezi le ezeket az értékeket:

$$\mathrm{T(x)} = \frac{e^{\mathrm{x}}}{1 + e^{\mathrm{x}}} \quad (2)$$

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

$$\pi_{(x)} = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots \ldots + \beta_k x_k}} \quad (3)$$

Majd a log-likelihood függvény maximalizálásával becsüljük az együtthatókat. Brimacombe (2016) szerint a log-likelihood a bemeneti adatokra való illesztésének a mértéke, valamint azt mutatja, mennyire valószínű, hogy a modell előrejelzi a megfigyelt adatokat.

Wilkinson et al. (2022) szerint az esélyhányadosok (odds ratios) hasznosak a logisztikus regresszió eredményeinek az értelmezésében. Az esélyhányadosok segítségével könnyebben lehet értelmezni, hogy a különböző független változók milyen mértékben és irányban befolyásolják a nemteljesítési valószínűséget. A β_j együttható esélyhányadosa a következőképpen számítható ki:

$$\mathrm{OR_J} = \frac{\mathrm{P(y = 0|x_j + 1)}}{\mathrm{P(y = 1|x_j + 1)}} \div \frac{\mathrm{P(y = 0|x_j)}}{\mathrm{P(y = 1|x_j)}} = \mathrm{e^{\beta j}} \quad (4)$$

Az esélyhányadosok megmutatják, hogy egy egységnyi növekedés az x_j független változóban hogyan befolyásolja az esélyeket, annak érdekében, hogy a függő változó y értéke 1 legyen. Ha az esélyhányados nagyobb, mint 1, akkor a növekedés az x_j változóban fogja növeli annak az esélyét, hogy y=1 legyen. Ha az esélyhányados kisebb, mint 1, akkor a növekedés az x_j változóban csökkenteni fogja az esélyét annak, hogy y=1 legyen.

Döntési fa

Chang et al. (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.

Kim et al. (2019) szerint a döntési fa előnyei közé tartozik az 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 a túltanulás 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 együttes tanulási módszer, amelynek célja, hogy kezelje azt a problémát, hogy a döntési fák könnyen szabálytalan mintává válnak, ha több elágazásuk van. Ez a jelenség túllillesztéshez vezethet, mivel a variancia minden esetben magas lesz, még akkor is, ha alacsony az eltérés. 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.

A modellek teljesítmények összehasonlítására használt módszerek

A hitelminősítő algoritmusok teljesítményének összehasonlítására és értékelésére több módszer alkalmazható, amelyek segítenek meghatározni, hogy az egyes algoritmusok mennyire pontosak és megbízhatóak a nemteljesítő és jól teljesítő ügyfelek előrejelzésében. Az általam használt módszerek a klasszifikációs mátrix, az AUROC (AUC) és ROC görbe.

A klasszifikációs mátrix négy elemet tartalmaz, amelyek alapján mérhető az algoritmus teljesítménye. Ez a négy elem a következő: az *igaz pozitív arány* (true positive rate) azoknak az ügyfeleknek az aránya, akik ténylegesen nemteljesítők és a modell helyesen azonosította őket; az *igaz negatív arány* (true negative rate) azoknak az ügyfeleknek az aránya, akik ténylegesen teljesítők és a modell helyesen azonosította őket; a *téves pozitív arány* (false positive rate) azoknak az ügyfeleknek az aránya, akik teljesítők, de a modell tévesen nemteljesítőnek minősítette őket; valamint a *téves negatív arány* (false negative rate) azoknak az ügyfeleknek az aránya, akik nemteljesítők, de a modell tévesen teljesítőnek minősítette őket.

A klasszifikációs mátrix alapján kiszámíthatóak a következő mutatók: a *pontosság* (accuracy), amely az összes helyesen osztályozott ügyfél aránya az összes ügyfélhez viszonyítva; az *érzékenységi mutató* (sensitivity), amely azt mutatja, hogy a pozitív osztályba tartozó megfigyelések mekkora százaléka van helyesen azonosítva. A *specificitás* (specificity) mutató, ami azt jelzi, hogy a negatív osztályba tartozó megfigyelések hány százalékban lettek helyesen azonosítva.

Shi et al. (2019) szerint az AUROC (AUC), vagyis a Receiver Operating Characteristic (ROC) görbe alatti terület, egy fontos mérőszám a klasszifikációs modellek teljesítményének értékelésére. A ROC görbe az algoritmus érzékenységét (sensitivity) ábrázolja a téves pozitív aránnyal különböző küszöbértékek mellett. A görbe alatti terület (AUC vagy AUROC) azt mutatja meg, hogy az algoritmus mennyire jól tud különbséget tenni a nemteljesítő és a teljesítő ügyfelek között. Ez az érték 0,5 és 1 között mozog, minél kisebb, annál rosszabbul teljesít a modell, minél közelebb van 1-hez, annál jobban teljesít a modell.

Eredmények

A változók közötti összefüggéseket vizsgáltam annak érdekében, hogy megállapítsam, hogy a nemteljesítés változót milyen tényezők befolyásolják. A khi-négyzet próbát használtam a minőségi változók esetén és a t-tesztet a mennyiségi változók esetén. A tesztek eredményei a mellékletben találhatók. A táblázatban fel van tüntetve a változók neve, a t, illetve a χ^2 statisztika értéke, valamint a szignifikanciaszint (p). Abban az esetben, ha a szignifikanciaszint kisebb 0,05-nél, kapcsolat mutatható ki az adott változó és a nemteljesítés között.

Tanuló és teszt adatbázis

Az adatbázist véletlenszerűen felosztottam tanuló és teszt adatbázisra. A felosztást követően az eredeti adatbázis 70%-a a tanuló adatbázis, 30%-a pedig a teszt adatbázis lett. A tanuló adatbázison történik a modellek becslése, a teljesít-

mények értékelése pedig a teszt adatbázison. A tanuló adatbázis 9228 jó ügyfelet, valamint 2733 rossz ügyfelet, a teszt adatbázis pedig 3954 jó ügyfelet és 1171 rossz ügyfelet tartalmaz.

Logisztikus regresszió becslése

Első lépésben lépésenkénti (stepwise) logisztikus regressziót becsültem a nemteljesítési valószínűség becslésére a tanuló adatbázison. A stepwise logisztikus regresszió lépésről lépésre döntést hoz arról, hogy mely változókat kell hozzáadni vagy eltávolítani a modellből (Estran et al. 2022). Használata lehetővé teszi a modell automatikus optimalizálását és egyszerűsítését anélkül, hogy manuálisan döntést kellene hozni minden egyes változóról.

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	1,578
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_ownershipJelzalog	-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_lim	-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_statusSource Verified	0,149	0,057	2,606	0,009	1,161
verification_statusVerified	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
purposeVacation	0,950	0,343	2,769	0,006	2,586

Forrás: Saját számítás

A minősítés (grade), a jelenleg használt aktív rulírozó hitelszámlák száma (num_actv_rev_tl), a lakástulajdon típusa (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 az 5. táblázat tartalmazza.

A teszt adatbázison teszteltem a logisztikus regresszió előrejelző képességét. Az osztályozási hatékonyságot a klasszifikációs mátrix alapján vizsgáltam. A 6. táblázat a klasszifikációs mátrixot tartalmazza, amely a tényleges és logisztikus regresszió által előrejelzett értékeket mutatja.

6. táblázat: Klasszifikációs mátrix: logisztikus regresszió

Előrejelzett	Tényleges	
	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	1515	140
Nemteljesítés (1)	2439	1031

Forrás: Saját számítás

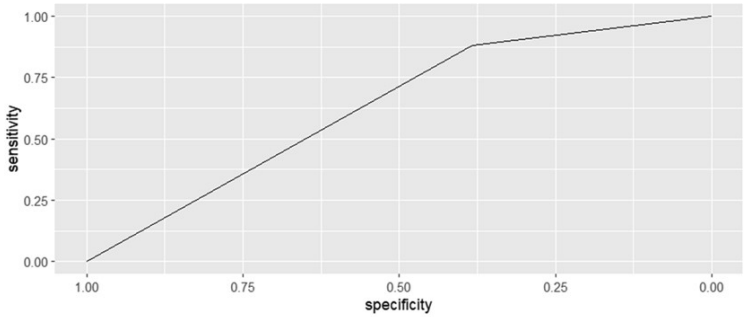
Az elemzés alapján a pontosság mutató értéke 0,4968, ami azt jelenti, hogy 49,68%-ban helyesen osztályozta az ügyfeleket a logisztikus regresszió. Az érzékenységi mutató 0,3832, ami azt mutatja, hogy a teljesítő osztályba tartozó ügyfelek 38,32%-a helyesen lett azonosítva. A specificitás mutató értéke 0,8804, ami azt jelzi, hogy a nemteljesítő ügyfelek 88,04%-a lett helyesen azonosítva. A pozitív előrejelzési érték 91,54%, míg a negatív előrejelzési érték 29,71%.

A ROC görbe grafikusan ábrázolja az osztályozó modell érzékenységét és specificitását különböző küszöbértékek mellett. A 2. ábra a ROC görbét szemlélteti a logisztikus regresszió esetén. Az AUC, vagyis a görbe alatti terület nagysága 0,6318, tehát a modell egy közepesen jó osztályozásba sorolható.

Ezután a WoE transzformált[2] adatok segítségével is becsültem a logisztikus regressziót. Stepwise logisztikus regressziót használva megállapítottam, hogy a hitel összege (loan_amnt_woe), a futamidő (term_woe), az al-minősítés (sub_

[2] A WoE-transzformáció lényege, hogy az adott adathalmazban lévő független változókat csoportokba rendezze a függő változó alapján, ami elősegíti az elemzést és a modellezést (Siddiqi 2012). A WoE-értékek azt mutatják meg, hogy az adott csoporthoz tartozó megfigyelések mennyire hordoznak információt a függő változó valószínűségének szempontjából. Minél magasabb a WoE-érték, annál jelentősebb és fontosabb információval rendelkezik a csoport az előrejelzések esetén.

grade_woe), a munkaviszony hossza (emp_length_woe), a lakástulajdon típusa (home_ownership_woe), az éves jövedelem (annual_inc_woe), a jövedelem ellenőrzési állapota (verification_status_woe), a lakhely (addr_state_woe), az összes adósságkötelezettség 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 a nemteljesítés előrejelzésében. Az eredmények a 7. táblázatban láthatóak.



Forrás: Saját szerkesztés

2. ábra: ROC görbe: logisztikus regresszió

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

Változó	Együttható	Std. hiba	Z érték	Pr(>|z|)
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

A 8. táblázatban a WoE-transzformációval becsült logisztikus regresszió teljesítménye látható a klasszifikációs mátrix alapján.

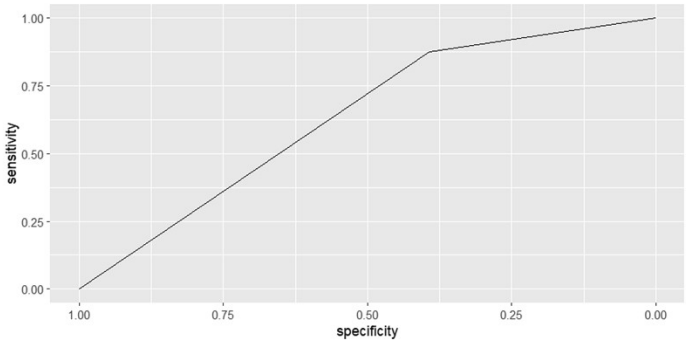
8. táblázat: Klasszifikációs mátrix: logisztikus regresszió WoE-transzformációval

Előrejelzett	Tényleges	
	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	1552	147
Nemteljesítés (1)	2402	1024

Forrás: Saját számítás

Az elemzés alapján 50,26%-ban helyesen osztályozta az ügyfeleket a logisztikus regresszió WoE-transzformációval. A teljesítő ügyfelek 39,25%-a, míg a nemteljesítő ügyfelek 87,45%-a lett helyesen azonosítva a modell által. A pozitív előrejelzési érték 91,35%, míg a negatív előrejelzési érték 29,89%.

Az AUC értéke 0,6335, tehát a modell hasonlóan különbözteti meg a teljesítő és nemteljesítő ügyfeleket az adott adathalmazon, mint a nem transzformált adatokon becsült logisztikus regresszió. A ROC görbe a 3. ábrán látható.



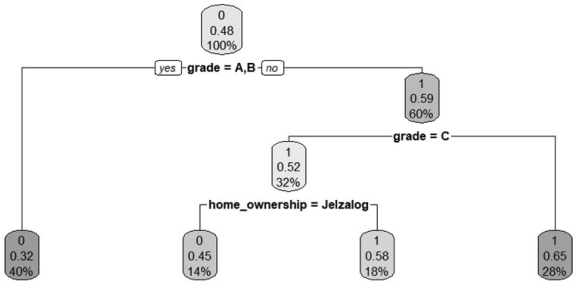
Forrás: Saját szerkesztés

3. ábra: ROC görbe: logisztikus regresszió WoE-transzformációval

Döntési fa

A döntési fát alul-mintavételezéssel használtam az első lépésben. Moscato et al. (2021) szerint az alul-mintavételezés 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. Létrehoztam az alul-mintavételezett tanuló adatbázist, amelyben 2951 jó ügyfél, valamint 2733 rossz ügyfél volt.

Meghatároztam az X-relative error értékét is, amely megmutatja, hogy a keresztvalidáció során mért elemzés szerint a fa mennyire teljesít jól az adott adathalmazon. Minél kisebb az érték, annál jobban általánosítható a kapott döntési fa. 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 lakástulajdon változó is figyelembe lesz véve. A CP-értékkel metszett döntési fa a 4. ábrán figyelhető meg.



Forrás: Saját szerkesztés

4. ábra: Metszett alul-mintavételezett döntési fa

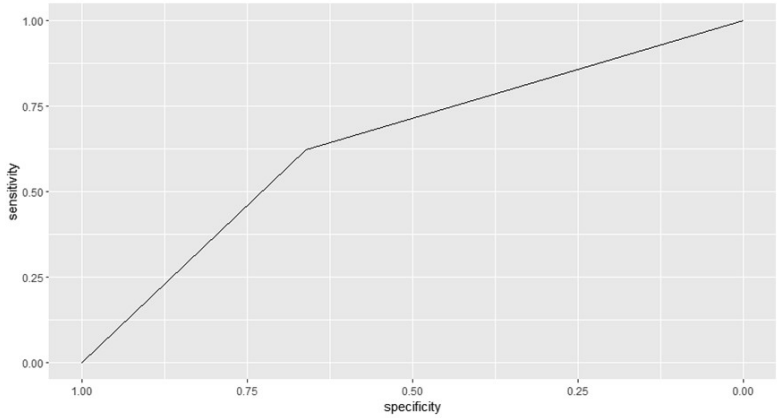
A teszt adatbázison teszteltem az alul-mintavételezett döntési fa előrejelző képességét klasszifikációs mátrixszal. Az eredmények a 9. táblázatban találhatók.

9. táblázat: Klasszifikációs mátrix: alul-mintavételezett döntési fa

Előrejelzett	**Tényleges**	
	Teljesítés (0)	**Nemteljesítés (1)**
Teljesítés (0)	2615	442
Nemteljesítés (1)	1339	729

Forrás: Saját szerkesztés

65,25%-ban helyesen osztályozta az ügyfeleket az alul-mintavételezett döntési fa. A teljesítő ügyfelek 62,25%-a, a nemteljesítő osztályba tartozó ügyfelek 66,14%-a lett helyesen azonosítva. A klasszifikációs mátrix után megrajzoltam a ROC görbét, amely az 5. ábrán látható. Az AUC értéke 0,642, tehát a modell egy közepesen jó osztályozásba sorolható.



Forrás: Saját szerkesztés

5. ábra: ROC görbe: döntési fa alul-mintavételezéssel

Shi et al. (2019) szerint a veszteségmátrix egy olyan technika, amely figyelembe veszi az osztályozási hibák költségeit a modell építése során. Ez igen hasznos egyensúlyhiányos adatkészletek esetén, ahol az egyes osztályok előrejelzési hibái különböző mértékű költségekkel járhatnak. Három különböző költségértéket 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-gyel.

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 nemteljesítés előrejelzésében. A veszteségmátrix segítségével meghatározott dön-

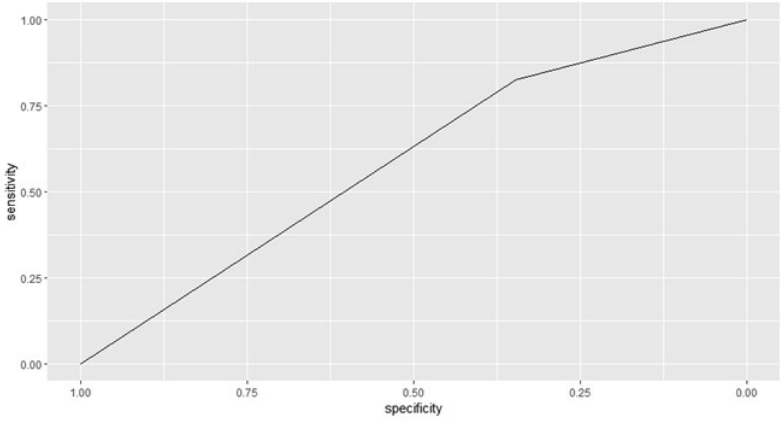
tési fa előrejelzési képességének vizsgálatához a klasszifikációs mátrixot használtam. A 10. táblázatban láthatóak az eredmények.

10. táblázat: Klasszifikációs mátrix: döntési fa a veszteségmátrix bevonásával

Előrejelzett	Tényleges	
	Teljesítés (0)	**Nemteljesítés (1)**
Teljesítés (0)	2615	442
Nemteljesítés (1)	1339	729

Forrás: Saját szerkesztés

Az elemzés alapján 65,25%-ban helyesen osztályozta az ügyfeleket a döntési fa. Az érzékenység mutató 0,6225, a specificitás mutató értéke 0,6614. A pozitív előrejelzési érték 35,25%, míg a negatív előrejelzési érték 85,54%. A 6. ábra a ROC görbét mutatja a döntési fa esetében. Az AUC, vagyis a görbe alatti terület 0,5866, tehát a modell egy gyenge osztályozásba sorolható.

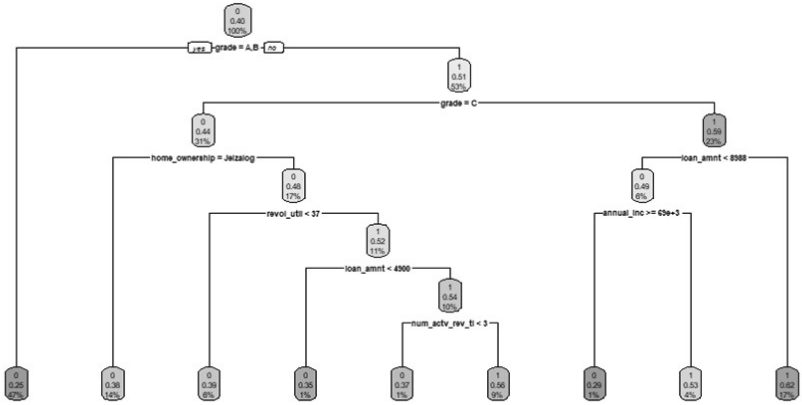


Forrás: Saját szerkesztés

6. ábra. ROC görbe: döntési fa veszteségmátrix bevonásával

Ezek után elkészítettem a döntési fát előzetes valószínűségek megadásával. Az előzetes valószínűségek megváltoztatásával becsült metszett döntési fa a 7. ábrán figyelhető meg. Az A vagy B csoportba sorolt ügyfelek esetén a nemteljesítési valószínűség 0,25. A nemteljesítési valószínűsége a D, E, F, G minősítésű

hiteleknek 0,53 abban az esetben, ha a hitel összege (loan_amnt) kisebb, mint 8988 USD, valamint az éves bevétel (annual_inc) kisebb, mint 68 500 USD.



Forrás: Saját szerkesztés

7. ábra: Döntési fa előzetes valószínűségek megváltoztatásával

A 11. táblázatban látható a döntési fa esetén a klasszifikációs mátrix.

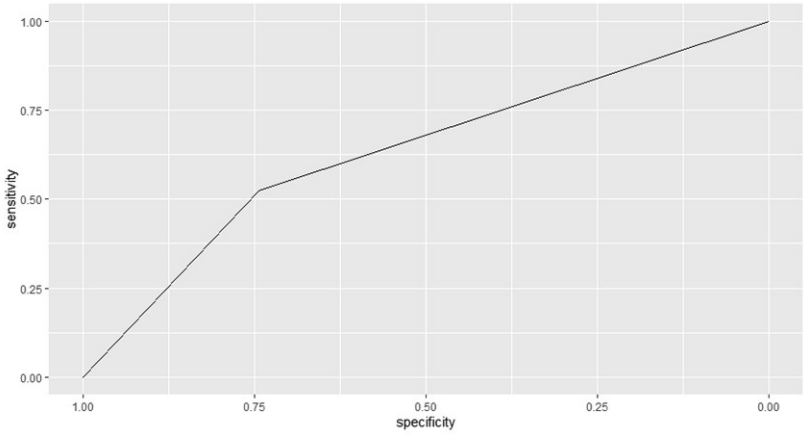
11. táblázat: Klasszifikációs mátrix: döntési fa előzetes valószínűségek megváltoztatásával

Előrejelzett	Tényleges	
	Teljesítés (0)	Nemteljesítés (1)
Teljesítés (0)	2939	558
Nemteljesítés (1)	1015	613

Forrás: Saját szerkesztés

A döntési fa 69,31%-ban helyesen osztályozta az ügyfeleket. A teljesítő ügyfelek 52,35%-a helyesen lett azonosítva. A specificitás mutató értéke 0,7433, ami azt jelzi, hogy a nemteljesítő ügyfelek 74,33%-a lett helyesen azonosítva. A pozitív előrejelzési érték 37,65%, míg a negatív előrejelzési érték 84,04%. A ROC görbe a 8. ábrán látható.

Az AUC értéke 0,6334, azaz 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 rosszabb eredményt, mint az alul-mintavételezés során.



Forrás: Saját szerkesztés

8. ábra: ROC görbe: döntési fa előzetes valószínűségek megváltoztatásával

Véletlen erdő

A véletlen erdő 500 döntési fa felhasználásával lett kialakítva, és az OOB (out-of-bag) becslés a hibaarányra 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 összesíti és ezáltal kap egy átlagos becslést a modell teljesítményére, anélkül, hogy külön validációs adathalmazt kellene létrehozni (Dzik-Walczak–Heba 2021). A teljesítő ügyfelek esetében a klasszifikációs hiba 11,81%, míg a nemteljesítő ügyfelek esetében 88,19% volt. A klasszifikációs mátrix a 12. táblázatban látható.

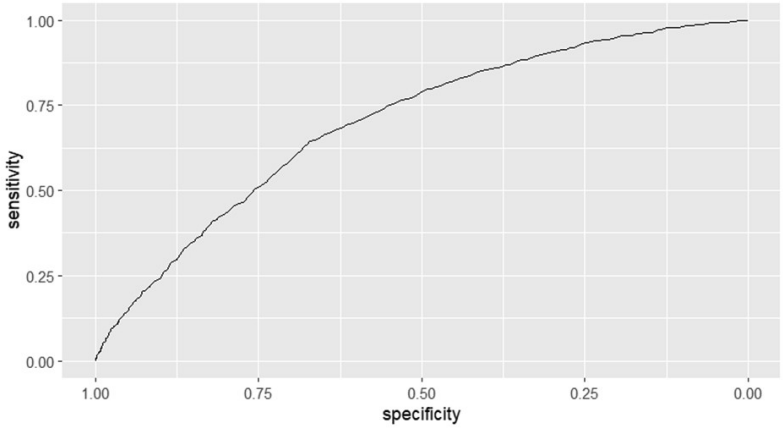
12. táblázat: Klasszifikációs mátrix: véletlen erdő

Előrejelzett	**Tényleges**	
	Teljesítés (0)	**Nemteljesítés (1)**
Teljesítés (0)	3886	1093
Nemteljesítés (1)	68	78

Forrás: Saját szerkesztés

A véletlen erdő 77,40%-ban helyesen osztályozta az ügyfeleket. A nemteljesítő ügyfelek 6,66%-a, a teljesítő ügyfelek 98,28%-a lett helyesen azonosítva.

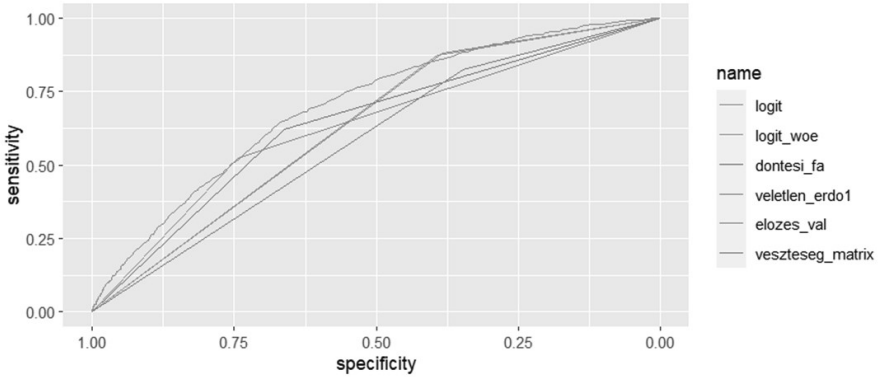
Az AUC értéke 0,7009, tehát a modell egy jó osztályozásba sorolható. A 9. ábra a ROC görbét mutatja a véletlen erdő esetén.



Forrás: Saját szerkesztés

9. ábra: ROC görbe: véletlen erdő

Annak érdekében, hogy jobban össze lehessen hasonlítani a modelleket, a 10. ábrán látható az összes becsült modell ROC görbéje.



Forrás: Saját szerkesztés

10. ábra: ROC görbék: összehasonlítás

A klasszifikációs mátrix, valamint a ROC görbe alapján arra a következtetésre jutottam, hogy a véletlen erdő a legjobban teljesítő modell az elemzettek közül. A 13. táblázat tartalmazza az összefoglaló statisztikákat.

13. táblázat: Összefoglaló táblázat

Modell neve	Pontosság	Érzékenység	Specificitás	PPV	NPV	AUC
Logisztikus regresszió	49,680%	38,320%	88,040%	91,540%	29,710%	0,631
Logisztikus-regresszió: WoE-transzformációval	50,260%	39,250%	87,450%	91,350%	29,890%	0,633
Alul-mintavételezett döntési fa	65,250%	62,250%	66,140%	35,250%	85,540%	0,642
Döntési fa a veszteségmátrix bevonásával	65,250%	62,250%	66,140%	35,250%	85,540%	0,586
Döntési fa előzetes valószínűségek megváltoztatásával	69,310%	52,350%	74,330%	37,650%	84,040%	0,633
Véletlen erdő	77,400%	6,660%	98,280%	53,450%	78,040%	0,700

Forrás: Saját számítás

Következtetések

A P2P-hitelezés egyre népszerűbbé válik mind a hitelfelvevők, mind a hitelezők számára. Az egyedi hitelkérelmek növekedésével egyre fontosabb lesz a hitelkockázat értékelése. A hagyományos módszereket használók egyre nagyobb kihívásokkal szembesülnek a változó gazdasági környezet és az adatok bősége miatt. Az új technológiák jó alternatívát kínálnak a hagyományos modellek mellett. A tanulmány azt vizsgálja, hogy az új technológiák képesek-e javítani a hitelminősítő intézmények hatékonyságát és eredményességét a hitelezési döntéshozatal folyamatában.

Klasszifikációs mátrix és ROC görbe, valamint AUROC segítségével vizsgáltam a logisztikus regresszió, döntési fa és véletlen erdő teljesítményét. Az eredmények alapján a véletlen erdő teljesített a legjobban.

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.

A tanulmány eredményeit összehasonlítottam a más cikkekben lévő eredményekkel. Az AUC legjobb értéke 0,700 a véletlen erdő esetén volt az elemzésben. A szakirodalomban, főként a modern gépi tanulási modelleknél gyakran találkozhatunk ennél magasabb AUC-értékekkel, akár 0,8 vagy magasabb is lehet egy jól optimalizált modell esetében. Ez arra utalhat, hogy a modell érzékenysége (sensitivity) és specificitása (specificity) nem a legoptimálisabb. A véletlen erdő modellnél mért 77,40%-os pontosság közelít a cikkek eredményeihez, de magasabb értékek is elérhetők megfelelő adatkezeléssel és modell-finomhangolással. A szakirodalomban gyakran találni olyan modelleket, amelyeknél a pontosság 80% fölött van, különösen kiegyensúlyozott adatok esetén. A logisztikus regressziónál és a döntési fáknál az érzékenység értékei viszonylag alacsonyak (38–66% körüliek), míg a specifikusság viszonylag magas. A szakirodalomban előfordul, hogy a kiegyensúlyozott adatok alapján felépített modelleknél mindkét érték magasabb, jelezve, hogy a modellek jobban tudják azonosítani mind a pozitív, mind a negatív osztályokat.

A kutatás javítható az adathalmaz méretének növelésével. Nagyobb adatbázis esetén a mintázatok jobb felismerésére képesek a modellek, ami javíthatja a modell teljesítményét. Az adatok kiegyensúlyozása, például felül-mintavételezés, alul-mintavételezés vagy szintetikus adatok generálása által javíthatja a modell érzékenységét. Fejlettebb modellek alkalmazása is segíthet a pontosság növelésén, a mélytanulási technikák, például a transzformer alapú modellek, sokszor jobban teljesítenek komplex mintázatok felismerésében. A hagyományos modellek fejlett algoritmusokkal való kombinálása is vezethet az eredmények pontosságának növeléséhez.

Irodalomjegyék

Ágoston, N. 2022a. Mesterséges intelligencia és gépi tanulási módszerek a vállalati fizetésképtelenség becslésére. *Statisztikai Szemle* 100(6), 584–609. https://doi.org/10.20311/stat2022.6.hu0584

Ágoston, N. 2022b. Külföldi csődelőrejelző módszerek szisztematikus irodalomelemzése. *Vezetéstudomány* 53(1), 69–89. https://doi.org/10.14267/VEZTUD.2022.01.06

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. https://doi.org/10.1051/e3sconf/202340905013

Alonso Robisco, A.–Carbo Martinez, J. M. 2022. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation* 8(1), 70. https://doi.org/10.1186/s40854-022-00366-1

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, https://doi.org/10.1016/j.stamet.2016.08.001

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. https://doi.org/10.3934/QFE.2022013

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. 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.

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

Fraisse, H.–Laporte, M. 2022. Return on investment on artificial intelligence: The case of bank capital requirement. *Journal of Banking & Finance* 138, 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. https://doi.org/10.1111/fima.12295

Kgoroeadira, 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. https://doi.org/10.1016/j.intfin.2023.101882

Milne, A.–Parboteeah, P. 2016. The Business Models and Economics of Peer-to-Peer Lending, *ECRI Papers* 11594, Centre for European Policy Studies.

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

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. https://doi.org/10.1016/j.techfore.2020.120462

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Shen, F.–Zhao, X.–Kou, G.–Alsaadi, F. E. 2021. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing* 98, 106852. https://doi.org/10.1016/j.asoc.2020.106852

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. https://doi.org/10.1016/j.frl.2019.03.033

Siddiqi, N. 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring* Vol. 3. John Wiley & Sons.

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

Stern, C.–Makinen, M.–Qian, Z. 2017. FinTechs in China–with a special focus on peer to peer lending. *Journal of Chinese Economic and Foreign Trade Studies* 10(3), 215–228. https://doi.org/10.1108/JCEFTS-06-2017-0015

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

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. https://doi.org/10.1016/j.ribaf.2022.101869

Wang, C.–Xiao, Z. 2022. A deep learning approach for credit scoring using feature embedded Transformer. *Applied Sciences* 12(21), 10995. https://doi.org/10.3390/app122110995

Wang, H.–Kou, G.–Peng, Y. 2021. Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society* 72(4), 923-934. https://doi.org/10.1080/01605682.2019.1705193

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. https://doi.org/10.1016/j.procs.2022.01.143

Wilkinson, J. D.–Mamas, M. A.–Kontopantelis, E. 2022. Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study. *Journal of Clinical Epidemiology* 152, 176–184. https://doi.org/10.1016/j.jclinepi.2022.09.009

Xia, Y.–Liu, C.–Li, Y.–Liu, N. 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* 78, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Zhao, C.–Li, M.–Wang, J.–Ma, S. 2021. The mechanism of credit risk contagion among internet P2P lending platforms based on a SEIR model with time-lag. *Research in International Business and Finance* 57, 101407. https://doi.org/10.1016/j.ribaf.2021.101407

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. https://doi.org/10.1016/j.physa.2019.122370

Melléklet: A változók közötti kapcsolat elemzése

Változó	t-statisztika	p
loan_amnt	-12,013	< 0,001
annual_income	7,831	< 0,001
Dti	-12,760	< 0,001
fico_range_low	22,491	< 0,001
fico_range_high	22,491	< 0,001
revol_util	-12,148	< 0,001
total_acc	2,348	0,018
total_rev_hi_lim	9,697	< 0,001
bc_open_to_buy	15,852	< 0,001
mo_sin_rcnt_rev_tl_op	8,072	< 0,001
total_cu_tl	-11,336	< 0,001
pub_rec_bankruptcies	-2,912	0,003
delinq_2yrs	-3,032	0,002
inq_last_6mths	-10,203	< 0,001
open_acc	-3,092	< 0,001
pub_rec	-1,321	0,186
acc_now_delinq	-1,065	0,287
total_cu_tl	1,570	0,116
chargeoff_within_12_mths	0,743	0,457
num_actv_rev_tl	-12,329	< 0,001
num_il_tl	0,215	0,831
num_op_rev_tl	-4,592	< 0,001
num_tl_op_past_12m	-8,799	< 0,001
Változó	**χ^2**	**p**
term (futamidő)	372,500	< 0,001
grade (minősítés)	1231,700	< 0,001
sub_grade (alminősítés)	1323,600	< 0,001
emp_length (régiség)	30,710	< 0,001
home_ownership (lakástulajdon)	120,721	< 0,001
verification_status (ellenőrzési állapot)	192,231	< 0,001
purpose (cél)	79,332	< 0,001
addr_state (lakhely)	103,413	< 0,001

Forrás: saját szerkesztés