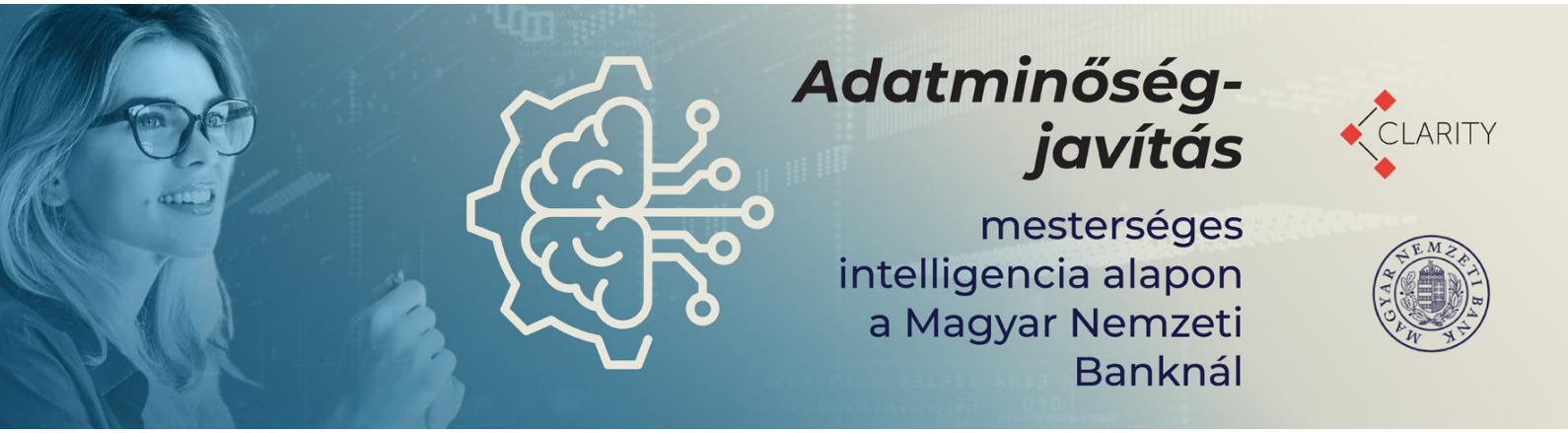



Szerző: Berndt Mihály

## Bevezetés

Napjaink gyorsan változó gazdasági környezetében a központi bankok szerepe a pénzügyi rendszer stabilitásának felügyelete miatt kiemelten fontos. Az MNB átfogóan elemzi a hazai pénzügyi szereplők tevékenységét, kockázatokat tár fel, jogszabályalkotásban vesz részt, felelőssége és felvigyázói feladatai vannak a pénzforgalom működését illetően.[1] Ehhez nem csak szimplán adatokra, hanem rendkívül jó minőségű adatokra van szükségük: a hitelek esetében például a granulált adatokat[2] tartalmazó adatbázis megléte kulcsfontosságú. Magyarországon a Hitelregiszter (HITREG) egy adatszolgáltatás, melyben a hitelintézetek havonta egyed szinten kötelesek beküldeni a vonatkozási időszakban kihelyezett és megszűnő hiteleiket egy 22 táblát és közel 500 dimenziót tartalmazó szerkezetben. Használatban van monetáris politikai eszközök kialakítására vonatkozó javaslatok, stratégiai és taktikai döntések előkészítése, bankok tevékenységének felügyelete és monitorozása során, továbbá számos más adatforrásokkal vetik össze, validációkra is használják.

A cikk összefoglalja a projektet, mely során a HITREG-ben gépi tanulás által korábban nem ismert anomáliákat detektáltunk az adatminőség növelése és az elemzők munkájának segítése érdekében.

[1] https://www.mnb.hu/penzugyi-stabilitas/az-mnb-feladatai-szerepe-a-penzugyi-stabilitasban

[2] Egyéni szintű információk. Az aggregációk néhány kiválasztott dimenzió mentén összesítettek akár több tízezer alacsonyabb szintű adatot, mely egy granuláris adatbázisban egyéni szinten rendelkezésre áll számos dimenzió mentén.

## Üzleti probléma

Az MNB és a hitelintézetek minden erőfeszítése ellenére az összegyűjtött adatokban vannak hibák, amelyek feltárása rendkívül nehéz analitikus módon, hiszen a már ismert összefüggések beépítésre kerültek a validációkba, újak feltárása hagyományos módszerekkel komplikált, az adatok mennyisége és heterogenitása miatt. A HITREG-ben történő anomáliák azonosítása ugyanakkor több okból is lényeges:

- Az adathalmazt számos osztály elemzője használja napi szinten a Bevezetésben említett feladatokra.
- Az adatszolgáltatók számára lényeges információkat jelenthet, akár adatkinyerési folyamatokban is feltárhat szisztematikus rendellenességeket.
- Új ellenőrzési szabályok generálását indukálhatja.
- Az algoritmus által feltárt mintázatok elemzése által mélyebb megértés alakulhat ki.

Az eddig nem ismert anomáliák feltárására az MNB Statisztikai igazgatóságával közösen egy gépi tanuláson alapuló megoldást fejlesztettünk ki. **Céljaink** a következők voltak:

- Szakértők által feldolgozható, **alacsony mennyiségű**, nagy mértékben **valóban anomáliának minősülő** rekord leválogatása, mivel a feladatra fordítható erőforrásuk számos esetben korlátos.
- Az MNB szakértői számára az általuk **meghatározott formában** bocsássuk rendelkezésre a gyanús megfigyeléseket, hangsúlyos az anomáliák gyors validálása.
- **Biztosítsunk** bizonyos szintű **magyarázatot** arra, hogy az algoritmus miért tartja a vonatkozó megfigyelést anomáliának, kiindulópontot adva a kiváltó okok elemzésére.

A projekt során az adathalmaz szűkítésre került azon hitelekre, melyek mögött fedezet áll. A megközelítés előnyei, hogy segít elérni a definiált célokat a következők miatt:

- hitelek esetében elengedhetetlen a fedezetek átfogó ellenőrzése;
- a megfigyelések értelmezhetők egy halmazként, összehasonlíthatók az anomáliák és a normál megfigyelések;
- hasonló dimenziók töltöttségét kényszerítik ki az ellenőrző szabályok;
- a rekordokat meghatározó karakterisztikák közösek;
- az említett jellemzők mérésére azonos táblákban lévő adatok vonhatók be az elemzés során.

## Megoldás és eredmények

A probléma megoldása során a gépi tanulás alkalmazásának legfőbb előnyei a következők:

- **Adatok mennyisége:** új mintázatok feltárása humánerőforrással, analitikus módon már nehéz, azonban gépi tanulási algoritmusokkal felfedhetők új összefüggések.
- **Rugalmasság**: a gépi tanulási algoritmusok képesek magas-dimenzionalitású, heterogén, összetett összefüggéseket tartalmazó adatokból információkat kinyerni.
- **Alkalmazkodóképesség:** a modell megfelelő monitorozása és szükség esetén újratanítása során anomáliák akkor is detektálhatók, ha az adatokban változások vagy nagy mértékű javítások következtek be.
- **Hatékonyság:** a pipeline során párhuzamos feldolgozásokat is alkalmaztunk.

A munka kiinduló fázisában két elterjedt megközelítést vizsgáltunk. Egyik esetben a rendelkezésre álló adatok felcímkézésre kerülnek: anomáliák (pl. korábban javításra került) és nem anomáliák, az algoritmus pedig megtanulja megkülönböztetni ezeket. Ez a megközelítés a következők miatt elvétésre került:

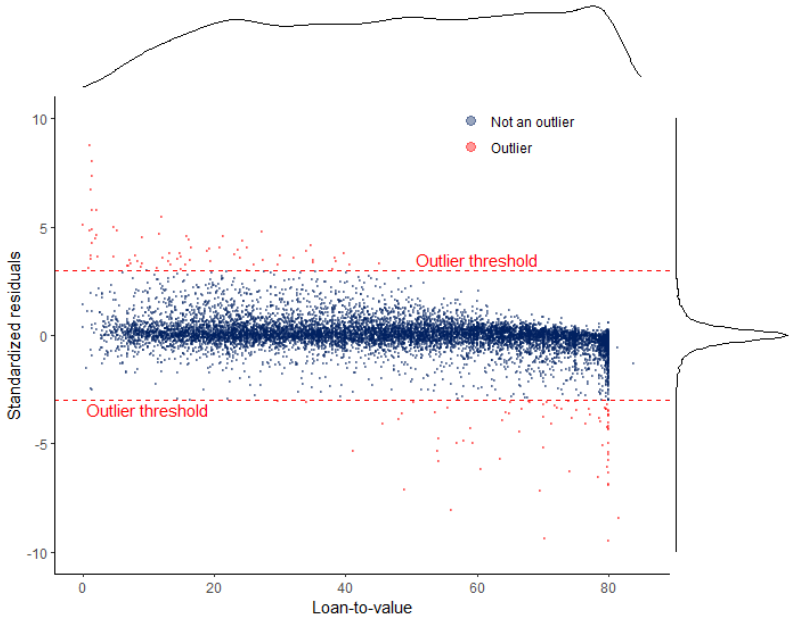
- ha szisztematikus hiba volt, akkor feltételezhető, hogy az adatszolgáltató javította rendszerében;
- ha az MNB kért javítást, akkor a hibát analitikus módon fedte fel, így ez jövőben is jó eséllyel felfedésre fog kerülni;
- a korábbi anomáliáktól eltérő karakterisztikával rendelkező hibákat csekély valószínűséggel képes felfedni.

Másik esetben a félig-felügyelt vagy nem felügyelt algoritmusokat kísérleteztünk, melyeknél a ritka régiókban elhelyezkedő megfigyelések az anomáliák, a többi megfigyelés pedig normális. Ezen alternatívák a következő hátrányaik miatt kerültek elvetésre:

- One-class SVM esetében a magas dimenzionalitás és rekordszám miatt skálázhatóság hiánya;
- hyperparaméterek beállítása Isolation forest és klaszterezések esetében jelentős mértékű ráfordítást igényel a HITREG csapat szakértőitől, mely nem lett volna hatékony;
- az interpretáció limitációja.

Végül a következő – kevésbé elterjedt, egyedibb és jelen esetben hatékonyabb – megoldást valósítottuk meg. A HITREG csapat szakértői által kiválasztásra került egy változó, a Loan

To Value (LTV), mely a hitel szerződéses összegének az arányát mutatja a fedezet(ek) becsült értékéhez viszonyítva. Az LTV értéke függ a többi változó értékétől. A megközelítés lényege, hogy egy gépi tanulási algoritmus egy regressziós feladatot old meg: az **algoritmus** megtanulja, hogy a magyarázó változók függvényében hogyan alakul az LTV értéke, majd új beküldéseknél **becslést ad az LTV-re. Amennyiben a tévedés mértéke (residual) meghalad egy küszöbértéket, akkor a megfigyelés anomália-jelölt lesz.**



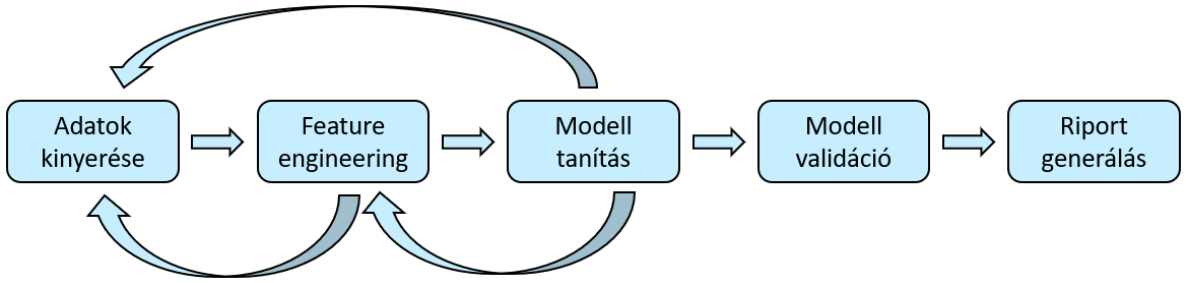

*1. ábra: Residual plot-on anomáliák megjelölése*

Megoldásunk előnyei a következők:

- **Kihasználja a szakértői tudást**: esszenciális, hogy a magyarázó változók függvényében hogyan alakul az LTV, a struktúrától eltérők anomália jelöltek.
- **Komplex kapcsolatokat képes felfedni:** a megtalált hibákat a hagyományos anomália-detektáló algoritmusok természetükből fakadóan nem biztos, hogy felfedték volna.
- **Paraméterezhető küszöbértéken alapuló detekció**: az anomáliák számossága könnyen a domain szakértők rendelkezésre álló erőforrásához igazítható.
- **Gördülékennyé teszi a döntéshozatalt:** egyértelmű és értelmezhető, hogy a várt mintázattól való eltérés alapján történik a hibák detektálása.
- **Flexibilitás:** a módszer rugalmasságot biztosít, más adathalmazokra is adaptálható szakértők bevonásával.

- **Nyomonkövethetőség a beavatkozás érdekében:** a modell-degradáció egyértelműen mérhető, így ha az adatokban változás történt, a modell újratanítható.
- **Interpretálhatóság:** mivel a megközelítés regresszión alapszik, értelmezhető, hogy mely magyarázó változók befolyásolták a vonatkozó rekord anomáliaként való megjelölését. Az interpretáció mértéke függ a modell komplexitásától.

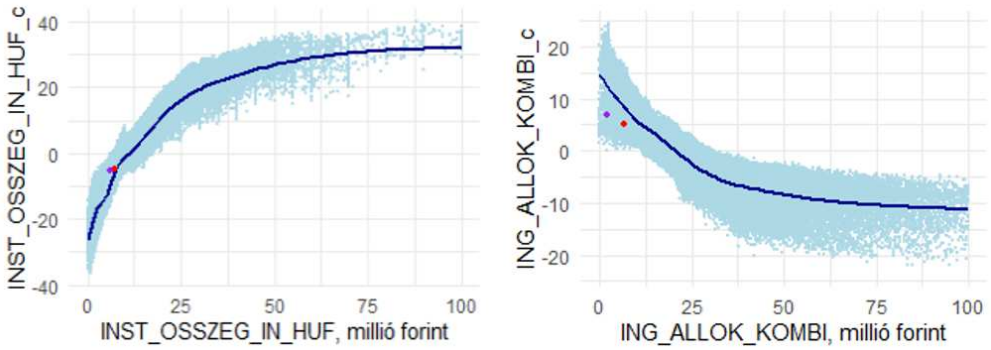
A kinyert adathalmazon alkalmazott modellezési pipeline meglehetősen hagyományosnak nevezhető, melyet felső szinten a következő ábra és az alatta olvasható felsorolás mutat be.




*2. ábra: Modellezési folyamat áttekintése*

A pipeline rövid, némileg technikaibb jellegű áttekintése a Mellékletben, a vonatkozó matematikai háttér ismertetése pedig a munkánkból készült hivatalos MNB Occasional Paper-ben olvasható.[3]

Fő eredménytermékeként egy Excel fájlt volt szükséges előállítani, melyben adatszolgáltatónként külön munkalapokra listáztuk a leválogatott anomáliajelölteket, a modell számára fontos változók értékeivel, azok adott rekordon történő becsléséhez való hozzájárulásukkal.




*3. ábra: Változók egyedi hozzájárulása LOESS simítással*

[3] https://www.mnb.hu/letoltes/mnb-op-148-final.pdf

Az ábrák X tengelyén az adott változó értékeinek egy szelete látható, az Y tengelyen pedig a becsléshez való hozzájárulás mértéke. A két kiválasztott változó hányadosa elméleti szinten az LTV – megjegyezve, hogy a jobb oldali egy származtatott változó. Látható, hogy a számláló növelésével nő az LTV becsült értékéhez való hozzájárulás, míg a nevező növelésével csökken, melyet megtanult a modell. A színes pontok két tetszőlegesen kiválasztott anomália-jelöltet mutatnak; megfigyelhető, hogy pusztán egy-egy érték mentén nem tűnnek gyanús eseteknek.

## Konklúzió

A megoldás egy nagy volumenű adathalmazban ember által már nagyon nehezen kivitelezhető módon végez leválogatást, szakértői validációja mégis gördülékenyen megtörténhet, mivel kevés és jó anomália jelöltet szelektál.

A célváltozó kiválasztását elsősorban domain tudás kell hogy vezérelje, megjegyezve, hogy érdemes olyan változót választani, melyben minimális a hiányzó értékek száma, ugyanakkor a magyarázó változókkal komplex a kapcsolata. Akár több modell is tanítható eltérő célváltozókkal: az anomália jelölteket lehet súlyozni a vonatkozó modell jóságának függvényében.

A megoldás sikerességének köszönhetően követő projektként már befejező fázisban jár egy újabb jelentős HITREG részhalmaz elemzése, mely fundamentálisan a bemutatott módszerre támaszkodik és csupán speciális adatelőkészítési és feature engineering lépésekben tér el.

## Melléklet: A pipeline bemutatása

A gépi tanulási pipepline egy struktúrált folyamat, mely a következő lépésekből áll:

- **nyers adatok kinyerése** forrásrendszerekből, esetünkben Oracle adattárház
- **feature engineering**
    - hiányzó adatok kezelésénél becslő függvényekkel, konstansokkal és az xgboost sparsity-aware split fittingjével próbálkoztunk
    - kategórikus változók esetén domain tudás alapján értékeket kombináltunk, mely segítette a túltanulás megelőzését
    - algoritmus számára magas információ tartalommal bíró változókat generáltunk
- **modell illesztése és kiértékelése**
    - különböző veszteségfüggvényekkel kísérleteztünk, a Huber jó választásnak bizonyult, mely egy beállított küszöbértéket meghaladó tévedéseknél lineáris, így ezen megfigyelések – melyek outlier-ek – kevésbé befolyásolják az algoritmust a paraméterek tanulása során
    - az említett küszöbértéken túl az xgboost számos további fontos hyperparaméterrel[4] rendelkezik, mely állítható a modellben, hangolásukat párhuzamosított Bayesian optimization módszerrel végeztük, legfőbb hozadéka a tanítási ciklusok futási idejének csökkentése.

[4] A hyperparaméterek olyan paraméterek, melyeket a modell nem tanul, hanem magát a tanulást befolyásolják.