Budapesti Műszaki és Gazdaságtudományi Egyetem
Önálló laboratórium (VIMIAL01)

Adatminőség kiértékelése

Általános adatminőség kiértékelő eszköz fejlesztése

Pósfay Benedek (CDJEOZ)
2024. május 30.

1 Bevezető

1.1 A projekt motivációja

A gépi tanulás területén az adatok minősége kiemelkedően fontos szerepet játszik. A jó minőségű adatok nemcsak a modellek pontosságát javítják, hanem megbízhatóbb és hitelesebb eredményeket is biztosítanak. Az adatminőség biztosítása érdekében szükség van olyan eszközökre, amelyek képesek az adatokat automatikusan ellenőrizni és kiértékelni.

A projektem célja egy általános adatminőség kiértékelő eszköz fejlesztése volt, amely Pythonban készült és a Great Expectations [1] könyvtárra épül. Az eszköz lehetővé teszi az adatforrások gyors és hatékony kiértékelését, segítve ezzel a gépi tanulási projektek adatminőségének javítását.

1.2 Az adatminőség fontossága a gépi tanulásban

A gépi tanulás során a modellek teljesítménye nagyban függ az adatok minőségétől. A rossz minőségű adatok számos problémát okozhatnak, többek között:

- **Zaj és hibás adatok**: A zajos és hibás adatok torzíthatják a modellek tanulási folyamatát, ami pontatlan előrejelzésekhez vezethet.
- **Hiányzó adatok**: A hiányzó értékek kezelése különös figyelmet igényel, mivel ezek befolyásolhatják a modell teljesítményét és megbízhatóságát.
- **Kiugró értékek**: A kiugró értékek jelentősen befolyásolhatják a modell pontosságát, különösen akkor, ha ezek az értékek nem reprezentatívak az adathalmaz többi részére nézve.

1.3 Korábbi tapasztalatok

Korábbi gépi tanulás projektjeim során többször szembesültem az adatminőség fontosságával, legutóbbi projektem során a kanadai lakossági energiafogyasztás előrejelző modell fejlesztésekor. A projekt célja az volt, hogy óránkénti energiafogyasztási adatokat [2] felhasználva pontos előrejelzéseket készítsünk a következő órák energiafogyasztására. Az adathalmaz több mint 600.000 sorból és több, mint 20 jellemzőből állt, amelyek tartalmazták az időjárási adatokat és a házak metaadatait is.

Az adatminőség problémája akkor vált nyilvánvalóvá, amikor észrevettem, hogy az adatok kevesebb mint 0,1%-ánál a fogyasztási adatok összezsúfolódtak egy órányi mérési adatba. Ez a hiba hatalmas kiugró értékeket eredményezett, amelyek jelentősen torzították a modell előrejelzéseit. A kiugró értékek miatt a modell nem tudott pontosan tanulni a valódi fogyasztási mintákról, ami pontatlan előrejelzéseket eredményezett.

A hibás adatokat kijavítva és a zajt csökkentve a modell pontossága jelentősen javult. A korrekciók után a modell pontossága több mint 50%-kal nőtt, ami megerősítette, hogy az adatminőség közvetlen hatással van a gépi tanulási modellek teljesítményére. A pontosabb adatok lehetővé tették a modell számára, hogy megbízhatóbb és hitelesebb előrejelzéseket készítsen, csökkentve a téves következtetések és rossz döntések kockázatát.

2 Kiindulás, fejlesztési irány meghatározása

Az eszköznek képesnek kell lennie különböző típusú adatforrások kiértékelésére, és rugalmasnak kell lennie a különböző elvárások és szabályok meghatározásában. Az eszköz fő jellemzői közé tartozik:

Automatikus validáció: Az eszköz automatikusan ellenőrzi az adatok különböző jellemzőit és szabályait, beleértve a numerikus, kategorikus, dátum és ID oszlopokat.

Illeszkedő szabályok: Az eszköz meghatározza az adatforrás különböző oszlopai alapján a megfelelő elvárásokat, így illeszkedve az adatforráshoz

Részletes jelentések: Az eszköz részletes jelentéseket generál HTML formátumban, amelyek tartalmazzák az adatminőségi tesztek eredményeit és statisztikáit.

Hibagyűjtés: Az eszköz összegyűjti azokat a sorokat, amelyek nem felelnek meg az elvárásoknak, és ezeket egy CSV fájlként exportálja a könnyebb elemzés érdekében.

2.1 Python és Great Expectations könyvtár használata

2.1.1 Miért Python?

Széles körben használt és támogatott: A Python az egyik legnépszerűbb programozási nyelv a gépi tanulás és adatfeldolgozás területén. Számos könyvtár és eszköz áll rendelkezésre, amelyek megkönnyítik az adatok kezelését és elemzését.

Korábbi tapasztalat gépi tanulás területen: Korábban több gépi tanulás témájú projektek során is Pythont használtam, adatelőkészítő kódrészletek újrafelhasználása céljából is a legalkalmasabb.

Gazdag ökoszisztéma: A Python gazdag ökoszisztémája számos könyvtárat tartalmaz, amelyek megkönnyítik az adatfeldolgozást, mint például a Pandas és a NumPy. Ezek a könyvtárak integrálhatók a projektbe, hogy még robosztusabbá tegyék az eszközt.

A Python kiválasztása azért is előnyös, mert széles körben támogatott a közösség által, és számos forrás áll rendelkezésre a problémák gyors megoldásához és a fejlesztési folyamat felgyorsításához.

2.1.2 Miért Great Expectations?

Specifikus célú könyvtár: A Great Expectations egy nyílt forráskódú könyvtár, amelyet kifejezetten adatminőség ellenőrzésére és dokumentálására fejlesztettek ki. Ez a könyvtár lehetővé teszi az adatminőség ellenőrzési szabályok egyszerű, de mégis robosztus és részletesen konfigurálható meghatározását és végrehajtását.

Automatikus validáció és jelentéskészítés: Az eszköz lehetővé teszi az adatforrások automatizált validációját és részletes jelentések generálását HTML formátumban. Ez segít a fejlesztés során felmerülő adatminőségbeli hibák gyors azonosításában és javításában.

Rugalmas és testreszabható: A könyvtár rugalmas és testreszabható, így könnyen adaptálható különböző típusú adatforrásokhoz és elvárásokhoz. Könnyedén bővíthető és kiegészíthető keretet ad az adatminőség kiértékeléséhez.

Könnyű integráció: A Great Expectations jól integrálható más Python alapú adatfeldolgozó eszközökkel és könyvtárakkal, mint például a Pandas. Ez lehetővé teszi az adatminőség ellenőrzési folyamat zökkenőmentes beillesztését a meglévő adatfeldolgozási munkafolyamatokba.

A Great Expectations választása azért volt különösen előnyös, mert a könyvtár kifejezetten az adatminőség ellenőrzésére készült, és számos beépített funkcióval rendelkezik, amelyek megkönnyítik az adatminőségi szabályok meghatározását, végrehajtását és dokumentálását. A könyvtár segítségével az adatminőségi problémák gyorsan és hatékonyan azonosíthatók és javíthatók.

3 Adatforrások

A projekt során több különböző adatforrást választottam ki az adatminőség kiértékelő eszköz tesztelésére és validálására. Ezek az adatforrások különböző típusú és méretű adathalmazokat képviseltek, amelyek lehetővé tették az eszköz rugalmasságának és hatékonyságának vizsgálatát.

Az első kiválasztott adatforrás a madridi ingatlanhirdetések adatbázisa [3] volt, amely nagyjából 21,000 sorból és több mint 40 jellemzőből állt. Ez az adatbázis változatos típusú adatokat tartalmazott, többek között az ingatlan hirdetési árát, alapterületét, a szobák számát és az ingatlan típusát (lakás, ház, iroda stb.). A nagy és változatos adatok ideálisak voltak az adatminőség kiértékelő eszköz tesztelésére, mivel lehetővé tették az eszköz különböző adatminőségi szabályainak és elvárásainak vizsgálatát, illetve előzetes vizsgálataim alapján bizonyos téren ismert hiányosságai vannak (üres oszlopok, nagy arányban hiányos jellemzők) és ezek vizsgálatával mérhető az értékelő eszköz minősége.

A második adatforrás a kanadai lakossági energiafogyasztás adatbázisa [2] volt a korábbi gépi tanulás témakörű projektemből, amely mintegy 600,000 sorból és több mint 20 jellemzőből állt. Az adatbázis tartalmazta az óránkénti energiafogyasztási adatokat, az időjárási adatokat, valamint a házak metaadatait. Ez az adatbázis részletes és nagy méretű volt, ami lehetővé tette az adatminőség kiértékelő eszköz skálázhatóságának és teljesítményének vizsgálatát nagy mennyiségű adat feldolgozása esetén.

A harmadik adatforrás egy kisebb mintát tartalmazott a madridi ingatlanpiac hirdetéseiből [5], körülbelül 900 sorral és 13 jellemzővel. Ez az adatbázis hasonló adatokat tartalmazott, mint az előző nagyobb adatbázis, így az eszköz segítségével összemérhető egy specifikus területen belüli két hasonló adatforrás minősége.

A negyedik adatforrás a melbourne-i ingatlanhirdetések adatbázisa [4] volt, amely több, mint 13,000 sorból és 20 jellemzőből állt. Az adatbázis tartalmazta az ingatlan hirdetési árát, alapterületét, a szobák számát és az ingatlan típusát, hasonlóan a madridi ingatlanhirdetések adatbázisaihoz. Ez az adatbázis is a hasonló területen belüli, de másik földrajzi elhelyezkedésű adatok tesztelését tette lehetővé, biztosítva, hogy az eszköz különböző földrajzi és piaci körülmények között is hatékonyan működjön.

Az adatforrások kiválasztása során figyelembe vettem a következő szempontokat:

- **Adatméret**: Különböző méretű adatbázisok választása a skálázhatóság és teljesítmény tesztelése érdekében.
- **Adattípusok**: Különböző típusú adatok (numerikus, kategorikus, dátum, ID) jelenléte, hogy az eszköz különböző adatminőségi szabályokat és elvárásokat teszteljen.
- **Földrajzi és piaci sokféleség**: Nemzetközi adatforrások választása, hogy biztosítsuk az eszköz rugalmasságát és alkalmazhatóságát különböző körülmények között.

Az adatforrások sokfélesége lehetővé tette az adatminőség kiértékelő eszköz alapos tesztelését és finomhangolását, biztosítva, hogy az eszköz különböző típusú és méretű adathalmazokkal is hatékonyan működjön. Ezek az adatbázisok kiváló alapot nyújtottak az eszköz teljesítményének és pontosságának mérésére, valamint az adatminőségi problémák azonosítására és kezelésére.

4 Kiértékelő algoritmus

Az adatminőség kiértékelő algoritmus célja, hogy automatikusan ellenőrizze az adatforrásokat különböző szempontok szerint, és egy 0-1 közötti értékkel jellemezze az adatminőségüket. Az algoritmus bemenetei közé tartozik az adatforrás (Pandas DataFrame), valamint néhány metaadat az adatforrásról, mint például az oszlopok típusai, a szigorúság mértéke és az exportálás részletességére vonatkozó paraméter.

Az algoritmus különböző típusú oszlopokra különböző elvárásokat fogalmaz meg. Ezek az elvárások különböző súlyokkal szerepelnek a végső értékelésben, attól függően, hogy mennyire fontosak az adott adathalmaz szempontjából. Az alábbiakban részletesen bemutatom az egyes oszloptípusokra vonatkozó elvárásokat és azok indoklását.

4.1 ID oszlopok

Az ID oszlopok egyedi azonosítókat tartalmaznak, amelyek az adathalmaz minden egyes sorát egyedileg azonosítják. Az ezekre vonatkozó elvárások a következők:

- **Ne legyen az érték NULL**: Az ID oszlopokban nem lehetnek hiányzó értékek, mivel minden sor egyedi azonosítóval kell, hogy rendelkezzen.
- **Ne legyen két egyforma érték**: Az ID oszlopban minden értéknek egyedinek kell lennie, hogy biztosítsa az adathalmaz sorainak egyértelmű azonosítását.
- **4 és 16 karakter között legyen a hossza**: Az ID értékek hosszának korlátozása biztosítja, hogy az azonosítók kezelhető méretűek legyenek.
- **Csak betűkből és számokból állhat**: Az ID értékeknek alfanumerikusnak kell lenniük, hogy elkerüljük az esetleges speciális karakterek okozta problémákat.

4.2 Numerikus oszlopok

A numerikus oszlopok számszerű adatokat tartalmaznak, amelyekre különböző statisztikai elvárások vonatkoznak:

- **Szám típusú értékek lehetnek**: Az oszlop minden értékének numerikusnak kell lennie.
- **Ne legyen az érték NULL**: A numerikus oszlopokban nem lehetnek hiányzó értékek, mivel ezek torzíthatják a statisztikai elemzéseket.
- **1 és 15 karakter között legyen a hossza**: Az értékek hosszának korlátozása segít biztosítani, hogy az adatok értelmezhető és kezelhető formátumban legyenek.
- **Értékek maximum 3.5 standard deviation távolságra lehetnek az átlagtól**: Az extrém kiugró értékek kiszűrésére, amelyek torzíthatják a statisztikai elemzéseket.
- **Értékek maximum 2.5 standard deviation távolságra lehetnek az átlagtól**: Egy szigorúbb szűrési szabály, hogy még inkább kizárjuk a potenciálisan hibás adatokat.
- **Értékek maximum 1.5 standard deviation távolságra lehetnek az átlagtól**: A legszigorúbb szűrési szabály a kiugró értékek kizárására, hogy biztosítsuk a nagy pontosságú adatelemzést.

4.3 Kategorikus oszlopok

A kategorikus oszlopok diszkrét kategóriákat tartalmaznak, amelyekre az alábbi elvárások vonatkoznak:

- **Ne legyen az érték NULL**: A kategorikus oszlopokban nem lehetnek hiányzó értékek, mivel ezek torzíthatják a kategóriákra vonatkozó elemzéseket.
- **1 és 15 karakter között legyen a hossza**: Az értékek hosszának korlátozása biztosítja, hogy a kategóriák kezelhető és értelmezhető formátumban legyenek.

4.4 Dátum oszlopok

A dátum oszlopok időbeli adatokat tartalmaznak, amelyekre az alábbi elvárás vonatkozik:

- **Legyenek dátum formátumúak**: Az oszlop minden értékének érvényes dátum formátumban kell lennie, hogy biztosítsuk az időbeli elemzések pontosságát és megbízhatóságát.

4.5 Szigorúsági paraméter és súlyozás

Az elvárások teljesülésének mértékét a szigorúsági paraméter határozza meg, amely 0 és 1 közötti érték lehet. Ez a paraméter meghatározza, hogy egy-egy elvárás az adatok legalább hány százalékára teljesüljön ahhoz, hogy az elvárást sikeresnek tekintsük. Az elvárások súlyozása attól függ, hogy mennyire fontosak az adott adathalmaz szempontjából. Például egy kiugró értékek szűrésére vonatkozó elvárás súlya nagyobb lehet, mint egy kisebb jelentőségű szabályé.

4.6 Kimeneti érték és exportálási lehetőségek

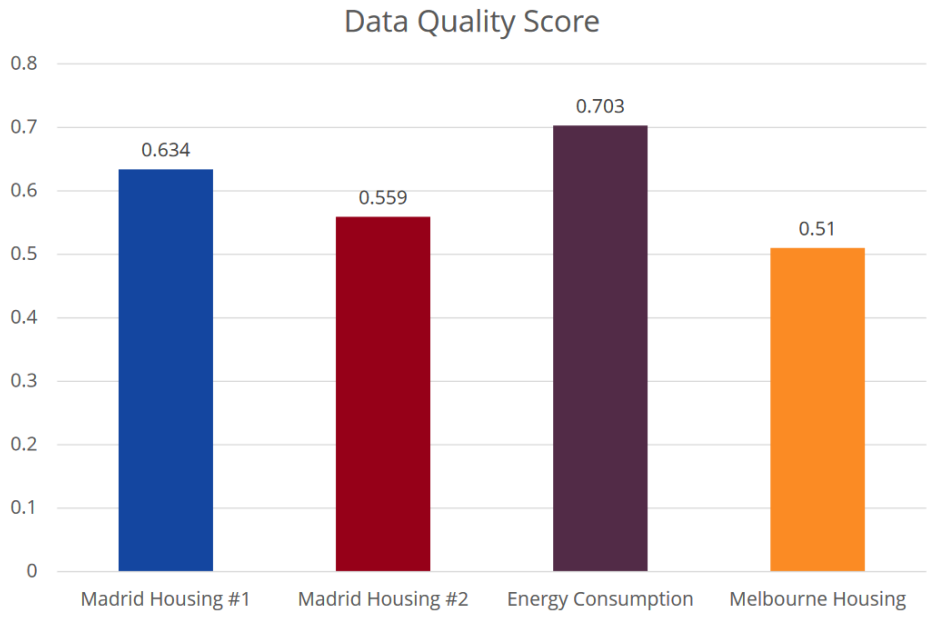
Az adatminőség kiértékelő algoritmus végső eredménye egy 0-1 közötti szám, amely az adott adathalmaz minőségét jellemzi. Emellett, ha az exportálás részletességére vonatkozó paraméter igaz, az algoritmus részletes jelentést generál HTML formátumban, amely tartalmazza az egyes futtatott tesztek eredményeit és statisztikáit. Az algoritmus egy hibagyűjtő funkciót is tartalmaz, amely összegyűjti azokat a sorokat, amelyek nem felelnek meg az elvárásoknak, megjelölve az egyes elvárások teljesülését is, és ezeket egy CSV fájlba exportálja a könnyebb elemzés érdekében. Az adatminőség kiértékelő eszköz tehát egy átfogó, rugalmas és hatékony megoldást nyújt az adatok minőségének ellenőrzésére és javítására, különböző típusú és méretű adathalmazok esetében.

5 Adatforrások értékelése

Az adatminőség kiértékelő algoritmus célja, hogy objektív módon értékelje az adatforrások minőségét. Ehhez az algoritmus az előző pontban bemutatott különböző elvárások alapján elemzi az adatokat, és egy összesített pontszámot generál, amely 0 és 1 között mozog. Minél magasabb ez a pontszám, annál jobb minőségűnek tekinthető az adott adatforrás. Az értékelés során különböző adatforrásokat vizsgáltam, hogy teszteljem az algoritmus hatékonyságát és megbízhatóságát.

5.1 Eredmények

A projekt során négy különböző adatforrást vizsgáltam meg, amikre az adatminőség kiértékelő algoritmus a következő eredményeket adta:



1. ábra: adatforrások értékelései

5.2 Eredmények elemzése

5.2.1 Madrid ingatlanhirdetések adatbázis (1. forrás)

A 0.634-es pontszám azt mutatja, hogy a madridi ingatlanhirdetések adatbázisa viszonylag jó minőségű, de van még hely a javításra. A legtöbb hibát az üres és nagyban hiányos oszlopok nagyobb aránya eredményezte.

5.2.2 Madrid ingatlanhirdetések adatbázis (2. forrás)

A kisebb minta alacsonyabb, 0.559-es pontszáma arra utal, hogy több adatminőségi problémát tartalmazott, például hiányzó értékeket és inkonzisztens adatokat. Ezeket a problémákat érdemes lenne részletesebben megvizsgálni és javítani.

5.2.3 Kanadai lakossági energiafogyasztás adatbázis

A 0.703-as pontszám viszonylag magas minőséget jelez. Az algoritmus azonosított néhány kiugró értéket, de ezek száma viszonylag alacsony volt. Az adatbázis általánosan megbízhatóbb és pontosabb adatokat tartalmazott, ami kedvezően befolyásolja a gépi tanulási modellek teljesítményét.

5.2.4 Melbourne ingatlanhirdetések adatbázis

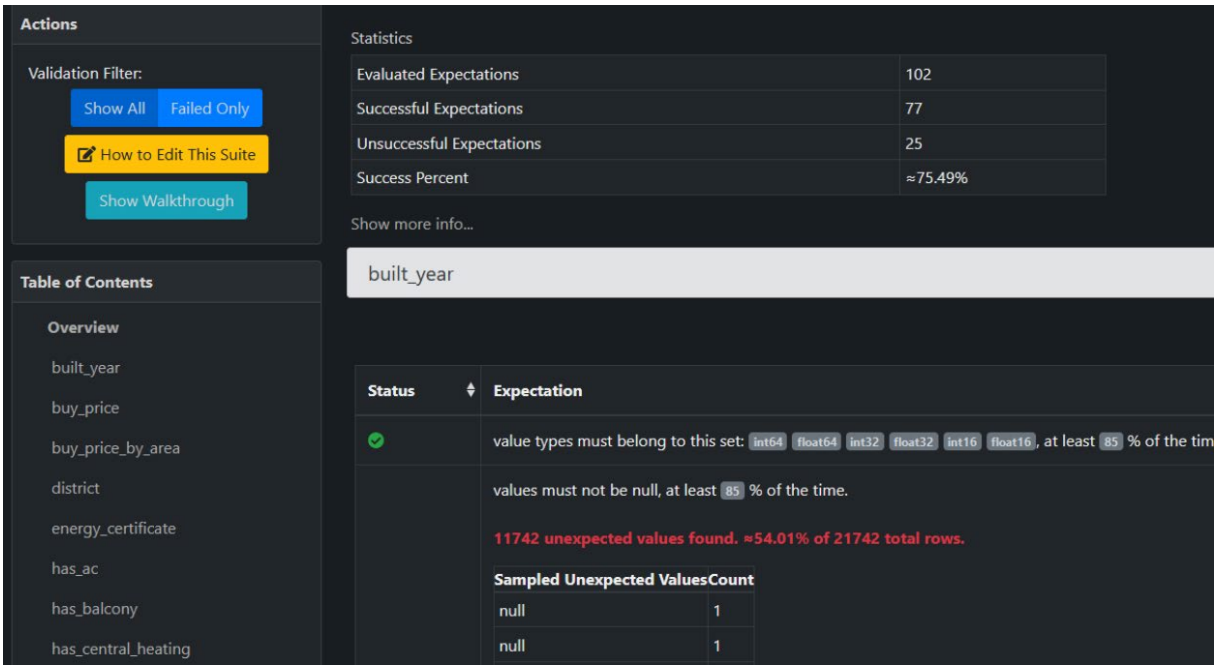
A 0.51-es érték viszonylag alacsony minőséget mutatott, ami arra utal, hogy az adatbázis jelentős adatminőségi problémákat tartalmazott a többi vizsgált adathalmazhoz képest. Az algoritmus számos hiányzó értéket, inkonzisztens adatot és kiugró értéket azonosított, amelyeket javítani kellene a pontosabb elemzések és előrejelzések érdekében.

6 Jelentések és hibaexport

Az adatminőség kiértékelő eszköz egyik fontos funkciója, hogy részletes jelentéseket generáljon és összegyűjtse azokat az adatokat, amelyek nem felelnek meg az elvárásoknak. Ezek a jelentések és a hibagyűjtő funkció segítenek a fejlesztőknek gyorsan azonosítani az adatminőségi problémákat és hatékonyan javítani azokat.

6.1 Jelentések generálása

A Great Expectations könyvtár használatával az eszköz részletes HTML formátumú jelentéseket készít, amelyek vizuálisan és strukturáltan mutatják be az adatminőség kiértékelésének eredményeit. A jelentések tartalmazzák az egyes tesztek eredményeit, statisztikáit és a különböző oszlopokkal kapcsolatos információkat.



2. ábra: általános jelentés részlet

6.2 *Jelentések tartalma*

Általános jelentés: Ez a jelentés összefoglalja az adatminőség kiértékelésének fő eredményeit. Tartalmazza a teljes adatforrás minőségét jellemző pontszámot, a sikeresen teljesült és a sikertelen elvárások arányát, valamint az adatforrás egészére vonatkozó statisztikákat.

Oszlopszintű statisztikák: Az egyes oszlopokra vonatkozó részletes statisztikák, mint például az érvényes és érvénytelen értékek száma és a hiányzó értékek aránya.

Elvárások eredményei: Az egyes elvárások teljesülésének eredményei, beleértve a sikeres és sikertelen sorok arányát. Az elvárások közé tartoznak az olyan szabályok, mint a numerikus oszlopok kiugró értékeinek ellenőrzése vagy a kategorikus oszlopok érvényességi vizsgálata.

Hibás értékek mintái: Példák azokra az értékekre, amelyek nem feleltek meg az elvárásoknak. Ezek a minták segítenek a fejlesztőknek jobban megérteni az adatminőségi problémákat és célzottan javítani azokat.

6.3 *Hibagyűjtő funkció*

Az eszköz hibagyűjtő funkciója összegyűjti azokat a sorokat, amelyek nem felelnek meg az egyes elvárásoknak, és ezeket egy CSV fájlba exportálja.

id	title	subtitle	raw_address	street_name	street_numbe	expectation_failed	house_type_id_length_betwee	has_central_heating_not_null	has_individual_heating_not_null
21216	Piso en ve	Los Ángele	Calle de la del M	Calle de la del M	65 -69	TRUE	FALSE	TRUE	TRUE
19700	Piso en ve	MoscardÃ³	Calle de Antonic	Calle de Antonio LÃ³pez		TRUE	FALSE	TRUE	TRUE
19474	Piso en ve	MoscardÃ³	Calle de Antonic	Calle de Antoni	115	TRUE	FALSE	FALSE	FALSE
19235	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18986	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18926	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18910	Estudio er	Cuatro Car	Calle Orden	Calle Orden		TRUE	TRUE	FALSE	FALSE
18771	Estudio er	Cuatro Car	Calle Orden, 6	Calle Orden	6	TRUE	TRUE	TRUE	TRUE
18730	Piso en ve	Bellas Vist	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	TRUE	TRUE
18676	Piso en ve	Cuzco-Cas	Calle de la Infan	Calle de la Infanta Mercedes, Mac		TRUE	FALSE	TRUE	TRUE
18512	DÃºplex en	Cuzco-Cas	Calle Pensamie	Calle Pensamie	27	TRUE	FALSE	TRUE	TRUE
18426	Estudio er	TetuÃ¡n, Madrid				TRUE	TRUE	TRUE	TRUE
18366	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviad	32	TRUE	FALSE	TRUE	TRUE
18306	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18260	Piso en ve	Cuzco-Cas	Calle Rosario Pi	Calle Rosario P	18	TRUE	FALSE	TRUE	TRUE
18251	Piso en ve	Cuzco-Cas	Calle de Rosaric	Calle de Rosari	18	TRUE	FALSE	TRUE	TRUE
18250	Piso en ve	Cuzco-Cas	Calle de Rosaric	Calle de Rosari	18	TRUE	FALSE	TRUE	TRUE

3. ábra: részletes hibajegyzék

Ez a funkció különösen hasznos a következők miatt:

Részletes hibajegyzék: Az exportált CSV fájl részletes hibajegyzéket tartalmaz, amelyben minden sor szerepel, amely nem teljesítette legalább az egyik elvárást. A fájl tartalmazza az eredeti adatokat és a hibás értékek konkrét elvárásait is.

Célzott javítás: A hibás sorok összegyűjtése lehetővé teszi a fejlesztők számára, hogy célzottan és hatékonyan javítsák az adatminőségi problémákat. Ahelyett, hogy az egész adathalmazt átvizsgálnák, a fejlesztők közvetlenül a problémás sorokra koncentrálhatnak.

Statisztikai elemzés: A hibás sorok statisztikai elemzése segíthet azonosítani a gyakori problémákat és mintákat, amelyek alapján javító intézkedéseket lehet hozni. Például, ha egy adott oszlopban sok hiányzó érték található, érdemes lehet megvizsgálni az adatgyűjtési folyamatot és javítani annak megbízhatóságát.

7 Jövőbeli fejlesztési irányok

A projekt keretében kifejlesztett általános adatminőség kiértékelő eszköz hatékonyan azonosította és jelentette az adatminőségi problémákat különböző típusú és méretű adathalmazok esetében. Azonban számos további fejlesztési irány létezik, amelyek tovább javíthatják az eszköz funkcionalitását és alkalmazhatóságát. Az alábbiakban részletesen bemutatom a jövőbeli fejlesztési irányokat.

7.1 Specifikusabb algoritmusok fejlesztése

Az egyik legfontosabb fejlesztési irány a specifikusabb algoritmusok kidolgozása, amelyek különböző felhasználási területekhez és adatforrás típusokhoz igazodnak. A jelenlegi általános algoritmus jól működik különböző adatforrások esetében, de bizonyos alkalmazások, mint például orvosi adatok, pénzügyi adatok vagy közösségi média adatok, speciális elvárásokat támaszthatnak az adatminőséggel szemben. Például:

- **Orvosi adatok**: Szigorúbb szabályok az adatok pontosságára és konzisztenciájára, mivel ezek közvetlenül befolyásolhatják a betegek kezelését.
- **Pénzügyi adatok**: Különös figyelem a hiányzó értékek és kiugró adatok kezelésére, mivel ezek jelentős hatással lehetnek a pénzügyi előrejelzésekre.
- **Közösségi média adatok**: Az adatok nagy volumenének és gyors változásának kezelése, valamint a szöveges adatok feldolgozása és minőségének ellenőrzése.

Az ilyen specifikus algoritmusok kifejlesztése lehetővé tenné az adatminőség kiértékelő eszköz alkalmazását szélesebb körben és nagyobb pontossággal az adott területre jellemző adatok esetében.

7.2 Oszlopok értékeinek Együttes vizsgálata

Jelenleg az algoritmus külön-külön vizsgálja az egyes oszlopokat, de a jövőben fontos lenne az oszlopok közötti kapcsolatok és korrelációk figyelembevétele is. Például egy ingatlanhirdetés adatbázisban a szobák száma és az ingatlan alapterülete közötti kapcsolat segíthet azonosítani az inkonzisztens adatokat. Az oszlopok együttes vizsgálata lehetővé tenné a komplexebb adatminőségi problémák felismerését és javítását.

7.3 YData Quality és más eszközök integrálása

A Great Expectations mellett érdemes lenne megvizsgálni más adatminőség ellenőrző eszközök, mint például a YData Quality, integrálását is. A YData Quality egy hasonló eszköz, amely további funkcionalitásokat kínálhat. Az eszközök kombinálása lehetővé tenné az adatminőség

ellenőrzési folyamatok további automatizálását és testreszabását, növelve ezzel az eszköz robosztusságát és alkalmazhatóságát.

7.4 Mesterséges intelligencia alapú megközelítés

A mesterséges intelligencia (MI) és a gépi tanulás integrálása az adatminőség kiértékelő eszközbe egy ígéretes jövőbeli fejlesztési irány. Az MI alapú megközelítések lehetővé tehetik a komplex minták és anomáliák azonosítását az adatokban, amelyek a hagyományos szabályalapú megközelítésekkel nehezen felismerhetők.

8 Összefoglalás és tanulságok

A projekt célja egy általános adatminőség kiértékelő eszköz fejlesztése volt, amely Pythonban készült és a Great Expectations könyvtárra épült. Az eszköz segítségével különböző típusú és méretű adatforrások minősége gyorsan és hatékonyan ellenőrizhető. Az eszköz különböző elvárásokat fogalmaz meg az adatforrások különböző típusaira vonatkozóan, és egy összesített pontszámot generál, amely az adatminőség mértékét jellemzi. A projekt során négy különböző adatforrást vizsgáltam meg, és az eredmények azt mutatták, hogy az eszköz hatékonyan azonosítja az adatminőségi problémákat.

A projekt során szerzett tapasztalatok és eredmények számos fontos tanulsággal szolgáltak. Először is, az adatminőség kiemelkedő fontossággal bír a gépi tanulási projektek sikeressége szempontjából. A jó minőségű adatok nemcsak a modellek pontosságát növelik, hanem megbízhatóbb és hitelesebb eredményeket is biztosítanak. Emellett a projekt rávilágított arra, hogy az adatminőség ellenőrzésének automatizálása és részletes jelentések generálása jelentős mértékben megkönnyíti a fejlesztők munkáját és javítja az adatok megbízhatóságát.

Hivatkozások

[1] Gong, A., Campbell, J., & Great Expectations. Great Expectations [Computer software]. https://github.com/great-expectations/great_expectations (lekérve: 2024. 05. 30.)

[2] Stephen Makonin. "HUE: The hourly usage of energy dataset for buildings in British Columbia", 2019, https://doi.org/10.1016/j.dib.2019.103744. (lekérve: 2023.12.12)

[3] Mirbek Toktogaraev. "Madrid real estate market", Kaggle, 2020, https://www.kaggle.com/datasets/mirbektoktogaraev/madrid-real-estate-market (lekérve: 2024. 05. 30.)

[4] Pino, Anthony. "Melbourne Housing Market", Kaggle, 2018, https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market (lekérve: 2024.05.30)

[5] Kanchana Karunarathna. "Madrid Idealista Property Listings [Data set]", Kaggle, 2024 https://doi.org/10.34740/KAGGLE/DSV/7764822 (lekérve: 2024. 05. 30.)