

Tartalomjegyzék

Bevezetés	1
A könyvről	2
Feltételezéseink rólad	3
A könyvben használt ikonok	4
A könyvön túl	5
Hol is kezdjük?	5

1. rész: ISMERKEDJÜNK MEG AZ ADATELEMZÉSSEL ÉS A PYTHONNAL..... 7

1. fejezet: Fedezzük fel, miért passzol a Python az adatelemzéshez..... 9

A Python mint programozási nyelv.....	10
Tekintsük át a Python – mint általános célú nyelv – számos használati módját	10
Értelmezzük a Pythont.....	12
Fordítsuk le a Pythont.....	12
Határozzuk meg, mi az az adattudomány.....	12
Hogyan jelent meg az adattudomány?.....	13
Vázzuk fel az adatfeldolgozók, adatelemzők alapvető kompetenciáit	13
Kapcsoljuk össze az adatfeldolgozást, a big datát és az MI-t	14
Alakítsuk ki az adatelemzési folyamatot	15
Ismerjük meg a Python szerepét az adatelemzésben	17
Nézzük meg, hogyan változik az adatfeldolgozók profílja.....	17
Használjunk egy univerzális, egyszerű és hatékony nyelvet	18
Gyorstalpaló a Python használatához	19
Töltsünk be adatokat	20
Tanítsunk be egy modellt	20
Jelenítsünk meg egy eredményt.....	21

2. fejezet: Ismerjük meg a Python képességeit..... 23

Használjuk a Pythont	24
Hozzájárulás az adatelemzési kutatásokhoz	24
Kóstoljunk bele!.....	25
Értsük meg, hogy kötelező behúzást használni.....	26
Használjuk a Jupyter Notebookot és a Google Colabot	26
Gyors prototípus-készítés és kísérletezés.....	27
A végrehajtási sebesség	28
Hatékony vizualizálás	30
Használjuk a Python-ökoszisztémát.....	31
Tegyük elérhetővé tudományos eszközöket a SciPy használatával	32
Végezzünk alapvető tudományos számításokat a NumPy használatával	32
Végezzünk adatelemzést a pandas használatával	32

Valóítsunk meg gépi tanulást a Scikit-learn használatával	33
Vessük bele magunkat a mély tanulásba a Keras és a TensorFlow használatával	33
Végezzünk elemzést hatékonyan az XGBoost használatával	34
Ábrázoljuk az adatokat a Matplotlib használatával	34
Készítsünk grafikonokat a NetworkX használatával	35
3. fejezet: Állítsuk be a Pythont adatelemzéshez	37
Használjuk az Anacondát	38
Használjuk a Jupyter Notebookot	38
Nyissuk meg az Anaconda-parancssort.....	39
Telepítsük az Anacondát Windowson.....	40
Telepítsük az Anacondát Linuxon	44
Telepítsük az Anacondát Mac OS X-en	45
Töltsük le az adathalmazokat és a példakódokat.....	47
Használjuk a Jupyter Notebookot	47
Határozzuk meg a kódtárat	48
Ismerjük meg a könyvben használt adathalmazokat	52
4. fejezet: Használjuk a Google Colabot.....	55
Határozzuk meg, mi a Google Colab	56
Nézzük meg, mire képes a Google Colab.....	56
Tekintsük át, miben tér el az online kódolás	58
Használjuk a helyi futtatókörnyezetes támogatást	59
Kezeljük jegyzetfüzeteket.....	60
Hozzunk létre egy új jegyzetfüzetet	60
Nyissunk meg meglévő jegyzetfüzeteket.....	61
Mentsünk el jegyzetfüzeteket	63
Töltsünk le jegyzetfüzeteket	66
Végezzünk el gyakori feladatokat	66
Hozzunk létre kódcellákat.....	66
Hozzunk létre szöveges cellákat.....	70
Hozzunk létre speciális cellákat	70
Szerkesszünk cellákat	71
Mozgassunk cellákat	71
Használjunk hardveres gyorsítást.....	71
Futtassuk a kódot	72
Tekintsük meg a jegyzetfüzetedet	73
Jelenítsük meg a tartalomjegyzéket.....	73
Kérjünk le információkat a jegyzetfüzetről	74
Ellenőrizzük a kód végrehajtását.....	74
Osszuk meg a jegyzetfüzetedet.....	74
Kérjünk segítséget.....	75
2. rész: ÁSSUK BELE MAGUNKAT AZ ADATOKBA.....	77
5. fejezet: Használjuk a Jupyter Notebookot.....	79
Használjuk a Jupyter Notebookot	80
Használjunk stílusokat.....	80
Kérjünk segítséget a Pythonhoz.....	82
Használjunk mágikus függvényeket	82

Fedezzük fel az objektumokat.....	84
Indítsuk újra a kernelt.....	85
Állítsunk vissza egy ellenőrzőpontot.....	86
Integráljunk multimédiás és grafikus tartalmakat.....	86
Ágyazzunk be diagramokat és más képeket	86
Töltsünk be példákat online weboldalokról.....	87
Töltsünk le online grafikus és multimédiás tartalmakat	87
6. fejezet: Dolgozzunk valódi adatokkal.....	91
Töltsünk be, streameljünk és mintavételezzünk adatokat	92
Töltsünk be kis mennyiségű adatot a memóriába	93
Streameljünk nagy mennyiségű adatot a memóriába.....	94
Kérjünk le képadatokat	95
Mintavételezzünk adatokat különböző módszerekkel.....	96
Férjünk hozzá strukturált listafájlok formájában tárolt adatokhoz	98
Olvassunk be szövegfájlból.....	99
Olvassunk be CSV tagolt formátumból.....	100
Olvassunk be Excel- és más Microsoft Office-fájlokat	102
Küldjünk adatokat strukturálatlan fájlok formájában.....	104
Kezeljünk relációs adatbázisokból származó adatokat.....	107
Kezeljünk NoSQL-adatbázisokból származó adatokat.....	108
Férjünk hozzá adatokhoz a weben.....	109
7. fejezet: Dolgozzunk fel adatokat.....	115
Váltogassunk a NumPy és a pandas között.....	116
Mikor érdemes a NumPy-t használni?.....	116
Mikor érdemes a pandast használni?.....	117
Érvényesítsük az adatokat	118
Derítsük ki, hogy mit rejtenek az adatok.....	118
Távolítsuk el a duplikátumokat	120
Készítsünk adattérképet és adattervet.....	121
Kezeljünk kategorikus változókat.....	123
Hozzunk létre kategorikus változókat	125
Nevezzük át szinteket	126
Vonjunk össze szinteket	127
Kezeljük a dátumokat az adatokban	128
Formázzunk dátum- és időértékeket.....	129
Használjuk a megfelelő időtranszformációt.....	129
Kezeljük a hiányzó adatokat	130
Keressük meg a hiányzó adatokat.....	131
Kódoljuk a hiányt	131
Rendeljünk hozzá hiányzó adatokat	133
Egy- és többirányú szeletelés: szűrjük és jelöljük ki az adatokat.....	134
Szeleteljünk sorokat	134
Szeleteljünk oszlopokat	135
Végezzünk többirányú szeletelést.....	136
Fűzzük össze és transzformáljunk adatokat	137
Adjunk hozzá új eseteket és változókat	137
Távolítsunk el adatokat	139
Rendezzük és keverjük össze az adatokat	139
Összesítsük az adatokat tetszőleges szinten.....	141

8. fejezet: Alakítsunk át adatokat	143
Használjuk a szózsákmodellt az adatok tokenekre bontásához.....	144
Ismerjük meg a szózsákmodellt.....	144
Rendezzük sorozatokba a szövegelemeket n-gramokkal	146
Valószínűsítsünk meg TF-IDF transzformációkat.....	148
Dolgozzunk gráfadatokkal	151
Ismerjük meg a szomszédsági mátrixot	151
Használjuk a NetworkX alapvető funkcióit	152
9. fejezet: Alkalmazzuk a tanultakat a gyakorlatban	155
Helyezzük kontextusba a problémákat és az adatokat.....	157
Értékeljük egy adattudományos problémát	157
Kutassunk megoldások után	158
Fogalmazzunk meg egy hipotézist	161
Készítsük elő az adatokat	162
A jellemzőalkotás művészete.....	162
Határozzuk meg, hogy mi az a jellemzőalkotás	162
Vonjunk össze változókat	163
Ismerjük meg a kategorizálást és a diszkretizálást	164
Használjunk indikátorváltozókat.....	165
Alakítsunk át eloszlásokat.....	165
Végezzünk műveleteket tömbökön	166
Használjunk vektorizálást	166
Végezzünk egyszerű aritmetikai műveleteket vektorokon és mátrixokon	167
Szorozzunk meg egy mátrixot egy vektorral.....	168
Szorozzunk össze mátrixokat	169
3. rész: VIZUALIZÁLJUNK ADATOKAT	171
10. fejezet: Végezzünk el egy Matplotlib-gyorstalpalót	173
Kezdjük egy grafikonnal	174
Definiáljuk a diagramot.....	174
Rajzoljunk több vonalat és diagramot	175
Mentsük lemezre a munkádat	176
Állítsuk be a tengelyt, a tengelyfeliratokat és a rácsokat.....	177
Kérjük le a tengelyeket	178
Formázzuk a tengelyeket	178
Adjunk hozzá rácsokat	180
Határozzuk meg a vonalak megjelenését.....	181
Használjunk vonalstílusokat	181
Használjunk színeket	183
Adjunk hozzá jelölőket	184
Használjunk címkeket, megjegyzéseket és jelmagyarázatokat.....	186
Adjunk hozzá címkeket.....	186
Fűzzünk megjegyzéseket a diagramhoz	187
Készítsünk jelmagyarázatot	188

11. fejezet: Vizualizáljuk az adatokat.....	191
Válasszuk ki a megfelelő diagramot	192
Készítsünk összehasonlításokat oszlopdiagramokkal.....	192
Jelenítsünk meg eloszlásokat hisztogramokkal.....	194
Ábrázoljunk csoportokat dobozdiagramokkal	195
Lássuk meg az adatmintázatokat pontdiagramokon	197
Készítsünk haladó szintű pontdiagramokat.....	198
Ábrázoljunk csoportokat.....	198
Jelenítsünk meg korrelációkat	200
Ábrázoljunk idősorokat	201
Ábrázoljuk az időt a tengelyeken	201
Ábrázoljunk időbeli trendeket	203
Ábrázoljunk földrajzi adatokat	205
Használjunk egy környezetet a Notebookban	206
Használjuk a Cartopy csomagot földrajzi adatok ábrázolásához ...	208
Kerüljük el az elavult könyvtárak használatát: a Basemap-eszköztár ...	211
Jelenítsünk meg gráfokat.....	212
Készítsünk irányítatlan gráfokat.....	212
Készítsünk irányított gráfokat	214
4. rész: SZERVEZZÜK ÁT AZ ADATOKAT	217
12. fejezet: Feszegessük a Python határait	219
Játsszunk a Scikit-learn könyvtárral	220
Ismerjük meg a Scikit-learn osztályait	220
Határozzuk meg az adattudományos alkalmazásokat.....	221
Használjunk transzformáló függvényeket	225
Láncoljunk egymás után becslőket	226
Transzformáljunk célokat.....	227
Állítsuk össze a jellemzőket	228
Kezeljük heterogén adatokat	229
Foglalkozzunk az időméréssel és a teljesítménnyel.....	231
Végezzünk teljesítményértékelést a timeit használatával.....	232
Használjuk a memóriaelemzőt.....	235
Futtassunk kódot párhuzamosan több magon	238
Használjunk többmagos párhuzamosságot	239
Szemléltessük a párhuzamos feldolgozást.....	241
13. fejezet: Tárjuk fel az adatelemzés rejtjelmeit.....	243
Az EDA megközelítés.....	244
Határozzuk meg numerikus adatok leíró statisztikáit	245
Mérjük a centrális tendenciát	247
Mérjük a szórásnégyzetet és a tartományt	248
Használjunk percentiliseket	249
Határozzuk meg a normalitás mérőszámait.....	250
Számoljunk össze kategorikus adatokat.....	252
Ismerjük meg a gyakoriságokat	253
Készítsünk kontingenciatáblázatokat	254
Készítsünk alkalmazott vizualizációt az EDA-hoz	255
Vizsgáljunk dobozdiagramokat	255

Végezzünk t-próbákat a dobozdiagramok megjelenítése után	257
Figyeljük meg a párhuzamos koordinátákat	258
Ábrázoljunk grafikusan eloszlásokat	259
Ábrázoljunk pontdiagramokat	261
Értsük meg a korrelációt	262
Használjunk kovarianciát és korrelációt	263
Használjunk nem paraméteres korrelációt	265
Vegyük számításba a kí-négyzetet táblázatoknál	266
Használjuk a Cramér-féle V-t	267
Módosítsuk az adateloszlásokat	267
Használjunk különböző statisztikai eloszlásokat	268
Készítsünk Z-pontszámú normalizálást	268
Transzformáljunk más nevezetes eloszlásokra	269
14. fejezet: Csökkentsük a dimenziószámot	271
Ismerkedjünk meg az SVD-vel	272
Nézzük meg, hogy lehet-e csökkenteni a dimenziószámot	273
Mérjük meg az SVD-vel, ami nem látható	276
Végezzünk faktoranalízist és főkomponens-analízist	276
Nézzük meg a pszichometriai modellt	277
Keressünk rejtett faktorokat	278
Használjunk komponenseket faktorok helyett	279
Érjük el a dimenziószám csökkentését	279
Nyomjuk össze az információkat a t-SNE-vel	280
Ismerkedjünk meg néhány alkalmazási területtel	283
Ismerjük fel arcokat a PCA segítségével	283
Nyeljünk ki témákat az NMF használatával	287
Ajánljunk filmeket	289
15. fejezet: Klaszterezés	293
Klaszterezünk a K-közép algoritmussal	294
Ismerjük meg a súlypontalapú algoritmusokat	296
Készítsünk egy példát képadatokkal	298
Keressünk optimális megoldásokat	299
Klaszterezünk big data adatokat	302
Végezzünk hierarchikus klaszterezést	304
Használjunk egy hierarchikus klasztermegoldást	305
Vizualizáljunk aggregáló klaszterezési megoldásokat	306
Tárjunk fel új csoportokat a DBScan használatával	308
16. fejezet: Észleljük a kiugró adatokat	311
Nézzük meg, hogy zajlik a kiugró értékek észlelése	312
Keressünk még több dolgot, ami elromolhat	313
Ismerjük meg, mik az anomáliák és az újszerű adatok	314
Vizsgáljunk meg egy egyszerű egyváltozós módszert	316
Használjuk ki a Gauss-eloszlást	318
Korrigáljuk a kiugró értékeket	319
Dolgozzunk ki egy többváltozós megközelítést	320
Használjunk főkomponens-analízist	321
Használjunk klaszteranalízist a kiugró értékek kiszűrésére	323
Automatizáljuk az észlelést izolációs erdővel	324

5. rész: TANULJUNK AZ ADATOKBÓL 327**17. fejezet: Ismerkedjünk meg négy egyszerű és hatékony algoritmussal..... 329**

Találjunk ki egy számot: lineáris regresszió.....	330
Írjuk le a lineáris modellek családját	331
Használjunk több változót	332
Ismerjük meg a korlátokat és a problémákat.....	334
Térjünk át a logisztikus regresszióra	335
Alkalmazzunk logisztikus regressziót	336
Tekintsük azt az esetet, amikor több osztály van	338
Egyszerűsítsük le a dolgokat annyira, mint a naiv Bayes-algoritmus.....	340
Ismerjük fel, hogy a naiv Bayes-algoritmus nem is olyan naiv	341
Jelezzük előre szövegek osztályozását	343
Használjunk lusta tanulást a legközelebbi szomszédokkal.....	345
Adjunk előrejelzést a szomszédok megfigyelése után.....	346
Válasszuk meg okosan a k paramétert.....	348

18. fejezet: Végezzünk keresztvalidációt, kiválasztást és optimalizálást..... 351

Mérlegeljük a modellillesztés problémáját	352
Ismerjük meg, mi az egyoldalúság és mi a változékonyság	354
Dolgozzunk ki stratégiát a modellek kiválasztásához	355
Válasszuk szét a gyakorló és a tesztalmodat.....	358
Végezzünk keresztvalidációt	361
Végezzünk keresztvalidációt k adatrészen.....	362
Vegyünk rétegzett mintát az összetett adatokból	363
Válasszuk ki a változókat profi módon	366
Válasszunk egyváltozós mérőszámok alapján	367
Alkalmazzunk előre és hátrafelé haladó kiválasztást	368
Turbózzuk fel a hiperparamétereket	369
Valósítsuk meg a rácskeresést.....	370
Próbáljuk ki a véletlenszerűsített keresést	374

19. fejezet: Növeljük a bonyolultságot lineáris és nemlineáris trükkökkel... 377

Használjunk nemlineáris transzformációkat.....	378
Végezzünk változótranszformációkat	379
Váltsunk ki kölcsönhatásokat változók között	381
Regularizáljunk lineáris modelleket.....	384
Támaszkodjunk a Ridge regresszióra (L_2)	385
Használjuk a Lassót (L_1)	386
Alkalmazzuk a regularizációt	387
Kombináljuk az L_1 -et és az L_2 -t: ElasticNet.....	387
Küzdjünk meg a big data adatokkal darabról darabra	388
Határozzuk meg, mikor túl sok az adat.....	389
Valósítsunk meg sztochasztikus gradiens csökkentést.....	389
Ismerkedjünk meg a szupport vektor gépekkel.....	393
Támaszkodjunk egy számítási módszerre.....	394
Rögzítsünk sok új paramétert.....	396
Osztályozzunk az SVC-vel.....	398
Nemlineárisra váltani egyszerű	404
Végezzünk regressziót az SVR-rel	406
Készítsünk sztochasztikus megoldást SVM-mel.....	408

Játsszunk neurális hálózatokkal	413
Ismerkedjünk meg a neurális hálózatokkal	414
Végezzünk osztályozást és regressziót neuronokkal	415
20. fejezet: Ismerjük meg a sokaság erejét	421
Kezdjünk egy egyszerű döntési fával	422
Ismerkedjünk meg a döntési fákkal	422
Készítsünk osztályozási fákat	425
Készítsünk regressziós fákat	428
Vesszünk el egy véletlen erdőben	429
Tegyük elérhetővé a gépi tanulást	430
Használjunk véletlen erdő osztályozót	433
Használjunk véletlen erdő regresszort	435
Optimalizáljunk egy véletlen erdőt	435
Gyorsítsunk az előrejelzéseken	437
Legyünk tisztában azzal, hogy a sok gyenge előrejelző nyer	437
Állítsunk be egy gradiens boosting osztályozót	439
Futtassunk egy gradiens boosting regresszort	440
Használjuk a GBM hiperparamétereit	440
Használjuk az XGBoost algoritmust	442
6. rész: TOP 10	445
21. fejezet: Tíz alapvető adatelemzési forrás	447
Tájékozódjunk a hírekről a Redditen	448
Adjuk meg a kezdőlökést a KDnuggetsszel	448
Kutassunk ingyenes tanulási források után a Quorán	449
Szerezzünk mögöttes információkat az Oracle AI & Data Science blogján	449
Nézzük meg a Data Science Central hatalmas forráslistáját	450
Fedezzünk fel új kezdő adatelemzési módszertanokat a Data Science 101-on	450
Szerezzük be a leghitelesebb forrásokat az Udacitynél	451
Kérjünk segítséget haladó szintű témákhoz a Conductricsnél	451
Szerezzük be a nyílt forráskódú adatelemzéssel kapcsolatos tudnivalókat a Springboardon	452
Célozzuk meg a fejlesztői forrásokat Jonathan Bowerrel	452
22. fejezet: Tíz adatelemzési kihívás	455
Távolítsuk el a személyazonosításra alkalmas adatokat	456
Alakítsunk ki biztonságos adatkezelési környezetet	457
Dolgozzunk több adatforrást használó problémákon	458
Csiszoljuk a túlillesztési stratégiánkat	459
Verekedjük át magunkat a MovieLens adathalmazon	459
Keressük meg a megfelelő adatforrást	460
Használjunk kézzel írt adatokat	461
Dolgozzunk képekkel	462
Az adatok életútjának meghatározása	463
Kezeljünk hatalmas gráfokat	464
Tárgymutató	467