

Киргизбаева Бибинур

Қазақ Ұлттық Аграрлық Зерттеу университеті

к.ф.-м.н., профессор

Тілеубай Әлішер Бақытбекұлы

Қазақ Ұлттық Аграрлық Зерттеу университеті

Қазақстан

Python БАҒДАРЛАМАЛАУ ОРТАСЫНДА МӘЛІМЕТТЕР ҚОРЫН ТАЗАЛАУ

Бүгінгі уақытта IT саласында жиі кездесетін мәселелердің бірі - қате деректер және оның алдын алу. Себебі үлкен деректер қоры (Big data) немесе деректер қоры - machine learning технологиясы, деректерді талдау, қолжазбаны немесе дауыстарды тану технологиялары үшін қолданылады, яғни деректер қоры тек сақтау үшін ғана емес, зерттеу мақсатында да қолданылады. Жоғарыда аталған технологиялар дұрыс жұмыс істеуі үшін деректер қоры қаншалықты таза және дұрыс құрылуы керек? Менің ойымша 95% - дан асауы қажет. Себебі деректер қорын пайдалана отырып зерттеген машина технологиясы өзінің шешім шығарады, яғни қате деректер қорын пайдаланған технология дұрыс жұмыс істемеуі әбден мүмкін. Әрбір деректер қоры ұяшығы - бұл машинаны оқыту үшін берілген есеп, сондықтан қазіргі уақытта деректер қорын тазартудың әртүрлі бағдарламалары жасалуда. Нақты және дұрыс толтырылған деректер қорына қол жеткізу үшін әртүрлі деректер қоры көріністерін шоғырландыру және қайталанатын ақпаратты өшіру қажет. Деректер қорын ғылыми зерттеуден басқа, талдау немесе сақтау үшін - ірі кәсіпорындар, компаниялар және мемлекеттік органдар да таза деректер қоры құрылуы қажет.

Жиі кездесетін деректер қоры мәселесі - ұяшықтардың бос жазылуы, дұрыс орнатылмаған тырнақшалар, тыныс белгілер, көшірмелер және т.б. Мұндай деректер көбінесе деректер қоры жүйесінің бұзылуына, машинаның дұрыс жұмыс істемеуіне және деректерді талдауда қате шешім шығаруына әкеледі.

Деректер қорын басқару үшін біз Python бағдарламалау тілінің Numpy, pandas, seaborn кітапханаларын қолданамыз. Бұл кітапханалар деректер қоры туралы негізгі ақпарат береді, график тұрғызады, бос бағандарды толтырады, барлық тыныс белгілері мен көшірмелерін алып тастайды.

Ақпаратты деректер қорында сақтамас бұрын тазарту керек - яғни бос ұяшықтарды толтыру, көшірмелерді және қажетсіз ақпараттар бағанын өшіру немесе өңдеу. Деректер қорын тазарту-бұл қате немесе бүлінген деректерді табу және өзгерістер енгізу, яғни өшіру немесе қатесін дұрыстау процесі. Процесс толық емес, қате, маңызды емес деректерді анықтайды, содан кейін "ластанған" деректерді ауыстырады, өзгертеді немесе жояды. Тазарту үшін біз Jupyter Notebook қолданамыз – бұл деректер қорын Python бағдарламалау тілін пайдалана отырып басқаруға мүмкіндік беретін веб әзірлеу ортасы. Веб әзірлеу ортасында сіз кітапханаларды орната отырып, деректер қорын манипуляциялау (деректер қорын тазарту) ғана емес ғылыми есептеу және машиналық оқыту саласындағы көптеген жұмыс процестерін жүргізе аласыз және веб интерфейсін баптай отырып өзгеде процестерді жүргізуге мүмкіндік аласыз. Мысалы ретінде біз әртүрлі тауарлар, артикулдар, сақтау мерзімі, тауарлар саны сақталатын қойма базасын аламыз.

Деректер қорымен жұмысты бастамас бұрын базаның өзін импорттап, жоғарыда аталған Pandas, numpy, seaborn кітапханаларын пайдалана отырып базаның көлемін және типін білуіңіз қажет. Бұл алдағы уақытта деректер қорымен жұмыс кезінде ориентациялау үшін. Деректер қоры data файлында, csv форматында сақталған. Shape және dtypes функциялары арқылы біз базаның көлемі мен типін білеміз.

```

17
18 # чтение данных
19 df = pd.read_csv('data.csv')
20
21 # shape and data types of the data
22 print(df.shape)
23 print(df.dtypes)

```

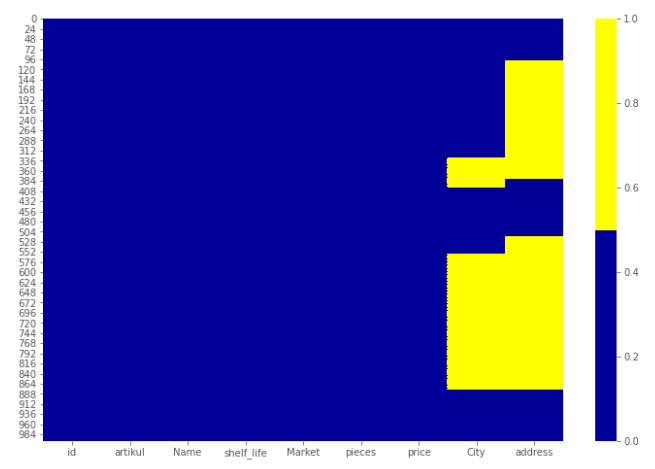
```

(1000, 9)
id                int64
artikul           int64
Name              object
shelf_life        object
Market            object
pieces            int64
price             object
City              object
address           float64
dtype: object

```

Сурет 1 – деректер көлемі және типі

Бұл код бізге деректер қоры 1000 жол, 9 бағаннан және int, object типтерінен тұратынын көрсетті. Енді біз деректер қорын "лас" мәліметтерді тазартуға кірісе аламыз. Жетіспейтін жолдармен жұмыс істеу-күрделі және кең таралған мәселелердің бірі. Жетіспейтін жолдарды анықтаудың үш әдісі бар: жетіспейтін жолдардың жылу картасын тұрғызу арқылы, жетіспейтін жолдардың пайызын шығару арқылы, жетіспейтін мәліметтердің гистограммасын тұрғызу арқылы. Біз Seaborn sns кітапханасының әдістерімен жылу картасын жасаймыз. Heatmap (df[cos].isnull(), с map=nt.color_palette (colors)), colors әдісімен бөлу үшін түстер береміз. Көрсетілген жылу картасы 30 жолдан тұратын жетіспейтін мәндерді көрсетеді. Белгілер көлденең, жолдарды жазу саны тігінен орналасқан.



Сурет 2 – жылу картасы.

Осылайша, біз жетіспейтін мәндерді визуалды түрді көріп ориентация жасай аламыз. Таңдалған жолдарды NULL немесе 0 мәндерімен толтыруға немесе бағанды жоюға болады, бұл қателіктердің алдын алмаған жағдайда деректер қорын сақтау барысында қате беруі мүмкін. Мысалы, ұяшықтардың, жолдардың анықталмауы немесе берілген ұяшық және кортеж элементтерің бос болуы білдіреді.

Жетіспейтін деректердің пайыздық тізімін жасау numru кітапханасы арқылы жүзеге асырылады. Numru көбінесе импортталған файлды математикалық есептеу үшін қолданады, бағандарға бөліп есептеп, орташа мәнді mean функциясын пайдаланамыз және 100-ге көбейтіп, жетіспейтін жолдардың пайызын біле аламыз.

Artikel, Name және т.б. бағандарда бос ұяшықтардың пайызы 0% көрсетеді, City – 39%, address – 64% және т. б. бұл тізім жетіспейтін деректердің жылу картасын толықтырады.

id	artikel	Name	shelf_life	Market	pieces	price	City	address	
1	2233305	Water - M	16-Aug-20	NASDAQ	18	78	JosefElv D	5647 Corben Trail	
2	8416916	Chicken -	11-May-20	NASDAQ	39	12	Jalaud	9199 Waubesa Court	
3	7467935	Beer - Sle	07-Jan-20	NYSE	76	7	Marechal	3 Erie Junction	
4	3721746	Soup - Car	18-Jul-20	NASDAQ	29	60	EhaydnDf	10813 Loomis Junction	
5	7447473	Hot Choc	15-Feb-20	NYSE	37	12	Pacatuba	4773 Ramsey Way	
6	7802477	Pancetta	23-Apr-20	NYSE	79	72		1052 Arapahoe Crossing	
7	6890453	Beans - Bl	17-Oct-20	NASDAQ	41	83		9222 Autumn Leaf Road	
8	8138052	Potatoes	02-May-20	NYSE	10	72		60 International Drive	
9	8558522	Water; Taj	02-Jan-20	NASDAQ	9	69		17 Mayer Point	
10	4199955	Corn - Mir	21-Sep-20	NASDAQ	85	19		90040 Anhalt Circle	
11	2078041	Worcester	16-Jun-20	NYSE	79	91		70 Dovetail Place	
12	1126043	Sprouts -	04-Aug-20	NYSE	32	64		6 Wayridge Place	

Сурет 3 – data.csvфайлы

Біз гистограмманы plot.bar функциясы арқылы құрамыз және біз бұл функцияны matplotlib кітапханасынан импорттаймыз. Сонымен қатар гистограмма жетіспейтін деректердің белгілерін анықтауға арналған жақсы визуализацияның бірі.

Жетіспейтін мәндерді қалай толтыруға болады. Қазіргі уақытта толтыру немесе жою туралы жалпы ортақ қолданатын шешімі жоқ, әрбір деректер қорына ерекше көзқарас қажет.Бос жолды қолдамай, бос ұяшыққа NULL мәнін қабылдайтын мәліметтер базасы бар, мұндай жағдайда біз кестені міндетті түрді

толтыруымыз керек. Егер бағанда бос ұяшықтары көп және ақпараттық құндылығы жоқ деректер сақталған болса, оны өшіріп тастауға болады. Мұндай бағандарды жоғарыда көрсетілген жылу картасынан немесе гистограммадан көруге болады. Мысалы city, address бағандары және т. б.

Ақпараттық құндылығы жоқ жазбаларды өшірудің көптеген жолдары бар, мысалы белгілі бір бағандарға шарт қою арқылы бағанды өшіре аламыз немесе жоғарыда есептеп алған пайызға шарт қою арқылы өшіруге болады. Яғни бергілі бір пайыздан асып кеткен жағдайда баған өшіріледі. Егер баған int типте болса, 0 немесе орташа мән қою арқылы бағанады өшірмей бос жолдарды толтыра аламыз.

Деректер қорын қате мәліметті ақпараттан тазарту қиынға соғады, себебі олар нақты деректерге сәйкес келуі мүмкін, сондықтан оларды анықтағаннан гөрі өшірген тиімді. Деректер типін анықтау үшін, сандық және категориялық сипаттамаларға сәйкес бөлу керек немесе нүктелік диаграммалар, кластер немесе Z - бағалауын құру қажет.

Median () функциясы медиананы табады және табылған мәнді price бағанының бос жолдарына толтырады. Медиан – реттелген сандар жиынтығының орташа мәні: деректердің жартысы осы мәннен төмен, ал жартысы жоғары.

```
Top = df[col].describe()['top']  
df[col] = df[col].fillna(top)
```

Коды арқылы бос жолдарды ең көп қайталанатын мәндермен толтырамыз. Describe () функциясы бірнеше мән береді, атап айтқанда орташа мән бұл RMS ауытқуы, бұл ең көп қайталанатын мән.

Осылайша бос жолдар толтырылады және біз csv форматында базаны импорттай аламыз. Бұл бос жолдардың қатесін болдырмайды.

321	320	1920620 Rabbit - Fr 19-Mar-20 NYSE	60 45.0	Beixiaoying	1 Westport Court
322	321	8188207 Onion - Dr 27-Jul-202 NASDAQ	60 3.0	Pinglin	99100 Homewood Street
323	322	7254198 Wine - Shi 11-Sep-20 NASDAQ	10 2.0	Trzebunia	96473 Anhalt Way
324	323	3362487 Slt - Indivi 09-Apr-20 NYSE	69 24.0	Haozigang	4 Dottie Avenue
325	324	3540558 Beef - Ter 16-Jan-20 NASDAQ	2 4.0	Boaco	64 American Ash Road
326	325	1498025 Cheese - C 13-Mar-20 NYSE	76 45.0	Kauit	2182 Bonner Center
327	326	1122817 Olives - Gi 15-Sep-20 NASDAQ	87 14.0	Ciawi	04786 Bluejay Court
328	327	1396045 Bread - M 24-Oct-20 NASDAQ	28 51.0	San Diego	075 Holy Cross Center
329	328	3376933 Chocolate 08-Mar-20 NYSE	62 51.0	San Diego	9 Rutledge Place
330	329	8127026 Bacardi Ra 01-Feb-20 NASDAQ	62 51.0	San Diego	31 Reindahl Lane
331	330	4838352 Wine - Stc 28-Jul-202 NYSE	29 51.0	San Diego	10 Mallard Alley
332	331	2552140 Wine - Nix 27-Apr-20 NYSE	99 51.0	San Diego	97 Dayton Terrace
333	332	6097144 Pork Casir 18-Jan-20 NASDAQ	58 51.0	San Diego	76103 Dunning Park
334	333	6713288 Sauce - So 31-Dec-20 NYSE	49 51.0	San Diego	6 Johnson Hill

Сурет 4 – нәтижесі

Толтырылған жолдар сары түспен белгіленген, бұл price және address бағандары. Price бағаны 51 медианасымен толтырылған, address бағаны ең көп қайталанатын Сан Диего мәні болып табылады.

Бұл мақалада біз лас деректердің түрлерін, олардың жұмысқа қалай әсер ететінін және таза деректердің қаншалықты маңызды екенін қарастырдық. Жоғарыда айтылғандай, көптеген нұсқалар бар қалай тазалау керек, бірақ жоқ жалпы алгоритм барлық мәліметтер базасы үшін бұл жұмысты қиындатады. Біз қолданған кітапханалар Numpy, Seaborn массивтерді өңдеу және деректерді талдау үшін қолданылады, бағытта біз деректерді тазарту, қате деректерді ауыстыру және бос жолдарды толтыру үшін қолдана аламыз.

Әдебиеттер

1. СиленД. ОсновыDataScience, BigData. Python и наука о данных / Д. Силен. М.: Питер, 2017.
2. Михнев И.П. Технологии Big Data и их применение в сфере современного высшего образования / И.П. Михнев, А.Д. Челнокова, А.Д. Реут / Развитие современного образования: от теории к практике: Материалы IV Междунар. науч.-практ. конф. (Чебоксары, 19 март 2018 г.) / Редкол.: О.Н. Широков – Чебоксары: ЦНС «Интерактив плюс», 2018.
3. Фрэнкс Б. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики / Б. Фрэнкс. М.: Альпина Диджитал, 2014.