Tipologia i cicle de vida de les dades

Pràctica 2

Mireia Calzada i Noemi Lorente

Pràctica 2

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de fins a 3 persones, o si preferiu, també podeu fer-ho de manera individual. Haureu de lliurar un sol fitxer amb l'enllaç Github (https://github.com) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu utilitzar aquests exemples com guia:

- Exemple: https://github.com/Bengis/nba-gap-cleaning
- Exemple complex (fitxer adjunt).

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació contiguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com).
Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al2009).
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).
- Predict Future Sales (https://www.kaggle.com/c/competitive-data-sciencepredict-future-sales/).

Els últims dos exemples corresponen a competicions actives a Kaggle de manera que, opcionalment, podrieu aprofitar el treball realitzat durant la pràctica per entrar en alguna d'aquestes competicions.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset forma part de la competició 'Predict Future Sales' de la plataforma Kaggle. Aquesta competició serveix de projecte final del curs online "How to win a data science competition", gestionat per Coursera, que permet aplicar i millorar les habilitats i competències d'un científic de dades.

El dataset conté l'històric de les dades de les vendes diàries de l'empresa **1C Company**, una de les companyies de programari russes més grans.

L'objectiu és predir les vendes de determinats productes i botigues durant el novembre de 2015. La predicció de vendes és una estratègia empresarial que ens permet:

- Millorar la gestió del capital humà per donar resposta a l'increment de vendes i donar millor servei al client.
- Detectar en quin moment de l'any es produeixen menys vendes per incentivar-les, per exemple creant promocions.
- Calcular la demanda dels productes, quina estacionalitat tenen, per a poder anticipar-nos en les comandes per proveir les botigues i no trencar estocs, és a dir, predir quan hem de fer una comanda.

Cal tenir en compte que la llista de botigues i productes varia lleugerament cada mes i gestionar aquestes situacions forma part del repte.

2. Integració i selecció de les dades d'interès a analitzar.

Abans de decidir quins camps utilitzarem i com els disposarem per a l'anàlisi, comencem fent un estudi de les dades de què disposem.

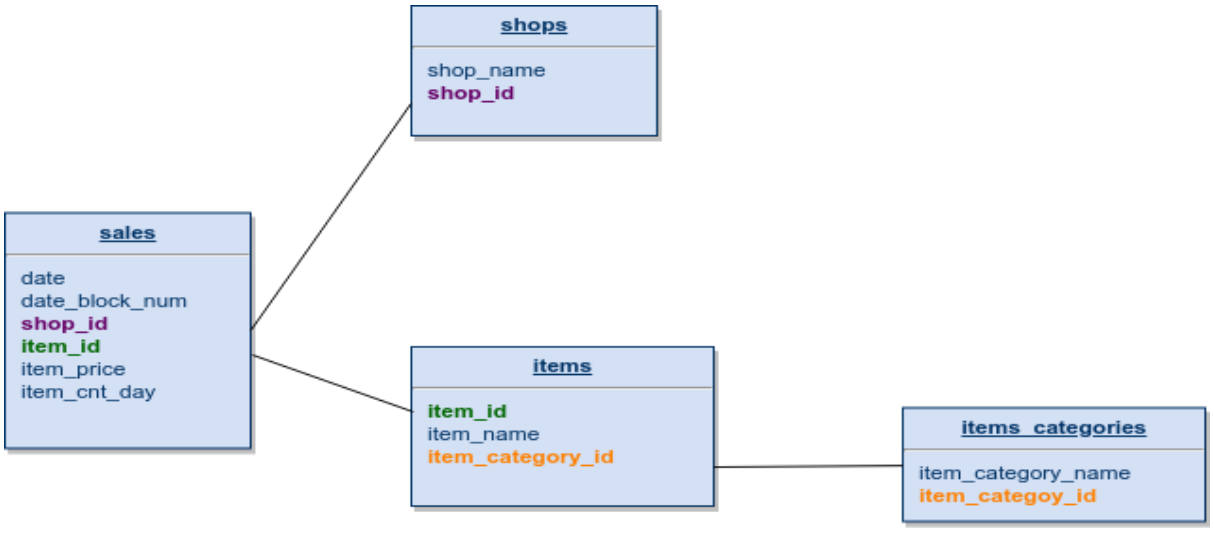
Descripció dels fitxers

- ***items.csv***: informació descriptiva sobre els articles / productes.
- ***item_categories.csv***: informació addicional sobre les categories dels productes.
- ***shops.csv***- informació descriptiva sobre les botigues.
- ***sales_train.csv****:* És el dataset d'entrenament i conté les dades històriques diàries des de gener de 2013 a octubre de 2015.
- ***test.csv***: És el dataset de prova. Cal predir les vendes d'aquestes botigues i productes per a novembre de 2015.
- ***sample_submission.csv***: un fitxer de mostra de la presentació de la predicció per producte.

Descripció dels camps de dades

Camp	Descripció
ID	Identificador que representa una tupla (un producte concret d'una botiga determinada) dins del conjunt de proves
shop_id	Identificador únic d'una botiga
shop_name	Nom de la botiga
item_id	Identificador únic d'un producte
item_name	Nom del producte
item_price	Preu de venda d'un producte
item_category_id	Identificador únic de la categoria d'un producte
item_category_name	Nom de la categoria del producte
item_cnt_day	Quantitat de productes venuts. L'objectiu és predir una quantitat mensual d'aquesta mesura
data	Data de la venda, en format dd / mm / aaaa
data_block_num	Número consecutiu que identifica els mesos. No s'inicialitza amb el canvi d'any. Per exemple, gener de 2013 és 0, febrer de 2013 és 1, ..., l'octubre de 2015 és 33.

Model E-R de les vendes per botiga



shops.csv

Comprovem que es tracta d'un dataset format per 60 registres amb 2 columnes:

- shop_id: identificador enter per la botiga
- shop_name: cadena de text amb el nom de la botiga.

```
> str(shops)
'data.frame':   60 obs. of  2 variables:
 $ shop_name: Factor w/ 60 levels "Адыгея ТЦ \"Мега\"",..: 57 59 1 2 3 4 5 6 7 8 ...
 $ shop_id  : int  0 1 2 3 4 5 6 7 8 9 ...
>
```

```
> head(shops)
                        shop_name shop_id
1  !Якутск Орджоникидзе, 56 фран       0
2  !Якутск ТЦ "Центральный" фран       1
3                Адыгея ТЦ "Мега"       2
4 Балашиха ТРК "Октябрь-Киномир"       3
5       Волжский ТЦ "Волга Молл"       4
6         Вологда ТРЦ "Мармелад"       5
>
```

Es tracta d'un conjunt de dades referents a vendes en botigues russes, de manera que tant els noms de les botigues com dels diferents productes estan en aquest idioma, la qual cosa ens dificulta la cerca de possibles registres duplicats.

De manera que traduïm els noms de les botigues a l'anglès i comprovem si hi ha dades repetides:

	shop_name	shop_id
9	Voronezh SEC-City Park "Grad"	8
10	Comerç sortint	9
11	Zhukovsky Str. Chkalov 39m?	10
12	Zhukovsky Str. Chkalov 39 m²	11
13	Botiga en línia Emergència	12
14	Centre comercial de Kazan "Behetle"	13
15	Centre comercial de Kazan "ParkHouse" II	14
16	Kaluga SEC "Segle XXI"	15

Observem que possiblement les botigues amb identificadors 10 i 11 facin referència al mateix establiment, doncs la descripció és pràcticament igual.

Ho tindrem en compte a l'hora de netejar les dades per tal d'agrupar les vendes diferents a aquestes dos botigues.

items_category.csv

Abans d'estudiar aquest fitxer de dades, traduïm els noms de les categories.

```
>
>
> str(item_categories_translated)
'data.frame':   84 obs. of  2 variables:
 $ PC...Auriculars...auriculars: Factor w/ 84 levels "Accessoris - PS2",..: 1 2 3 4 5 6 7 19 47 11 ...
 $ item_category_id            : int  0 1 2 3 4 5 6 7 8 9 ...
>
```

```
>
> head(item_categories_translated)
            item_category_name item_category_id
1 PC - Auriculars / auriculars                0
2              Accessoris - PS2               1
3              Accessoris - PS3               2
4              Accessoris - PS4               3
5              Accessoris - PSP               4
6           Accessoris - PSVita               5
>
```

Comprovem que es tracta d'un fitxer amb 84 registres, cadascun dels quals està format per 2 atributs:

- item_category_id: enter identificador de la categoria del producte
- item_category_name: cadena de text la descripció de la categoria.

items.csv

```
>
> str(items)
'data.frame':   22170 obs. of  3 variables:
 $ item_name       : Factor w/ 22170 levels "007 Legends [PS3, русская версия]",..: 9929 1122 9909 10558 132
53 16234 20048 20047 21590 21984 ...
 $ item_id         : int  0 1 2 3 4 5 6 7 8 9 ...
 $ item_category_id: int  40 76 40 40 40 40 40 40 40 40 ...
>
```

Comprovem en aquest cas que es tracta d'un fitxer amb 22170 registres formats per 3 variables que fan referència a:

- item_category_id: enter identificador de la categoria del producte
- item_id: enter identificador del propi producte
- item_name: cadena de text amb la descripció del producte

sales_train_v2.csv

Analitzem el fitxer de les vendes que farem servir per a la predicció.

```
> str(sales_train_v2)
'data.frame':   2935849 obs. of  6 variables:
 $ date          : Factor w/ 1034 levels "01.01.2013","01.01.2014",..: 35 69 137 171 477 307 35 103 341 69 .
..
 $ date_block_num: int  0 0 0 0 0 0 0 0 0 0 ...
 $ shop_id       : int  59 25 25 25 25 25 25 25 25 25 ...
 $ item_id       : int  22154 2552 2552 2554 2555 2564 2565 2572 2572 2573 ...
 $ item_price    : num  999 899 899 1709 1099 ...
 $ item_cnt_day  : num  1 1 -1 1 1 1 1 1 1 3 ...
>
```

En aquest cas tenim un conjunt de gairebé 3 milions de registres de 6 variables cadascun.
Observem que conté dades de vendes diàries en les diferents botigues (shop_id) i per als diferents productes (item_id). Inclou a més, la data de la venda (date), el preu (item_price) i el recompte d'unitats venudes (item_cnt_day) des de l'1 de gener del 2013 al 31 d'octubre de 2015.

El camp "data_block_num" mostra el mes en què es va produir la venda. Aquesta dada no la utilitzarem per a la predicció.
Tampoc utilitzarem el camp item_price, ja que no ens aporta informació extra ni necessària per a l'estudi.

Una observació interessant és que tenim quantitats negatives en la columna 'item_cnt_day'. Considerem que aquestes quantitats negatives són degudes a devolucions, de manera que les tindrem en compte a l'hora de predir les vendes.

Finalment, amb l'objectiu de predir la quantitat total de cada producte que es venen a cada botiga, a partir del fitxer sales_train_v2.csv, seleccionarem els camps:

- date
- shop_id
- item_id
- item_cnt_day

I aquest subconjunt de dades, el tractarem per tal de:

- Fusionar les vendes de les botigues 10 i 11
- Hi ha 3 botigues (9,20,36) que només estan obertes a l'octubre. Creiem que pot ser per necessitats de gran demanda. Mireia, creus que cal esborrar aquestes botigues?
- L'algorisme de predicció Prophet requereix com a paràmetre la data original. Al disposar d'observacions diàries de les mostres, la predicció de l'algorisme serà més precisa. El camp date és un string i té el format DD.MM.YYYY. Cal formatar-lo a tipus data (YYYY-MM-DD)

3. Neteja de les dades.
    a. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

- El fitxer sales_train no conté cap 0 en el camp quantitat venuda
- Comprovar valors nulls o fora de rang en item_category dins items
- Comprovar rangs items, items_category i shops en sales_train

    b. Identificació i tractament de valors extrems.

- boxplots de les diferents variables per separat: date, shop_id, item_id, item_cnt_day

4. Anàlisi de les dades.

Les series temporals són un conjunt d'observacions ordenades en el temps. Són valors dependents, cada valor successiu depèn de valors anteriors.

Volem crear un model matemàtic que dibuixi la relació de la variable item_cnt_day amb el temps i poder eliminar tant el soroll (tendència, estacionarietat i autocorrelació) com sigui possible, de manera que l'únic moviment de les dades sigui l'aleatorietat pura.

Podem definir la sèrie temporal additiva a partir de les seves components principals:

$$X = T \text{ (tendència)} + E \text{ (estacionarietat)} + A \text{ (aleatorietat)}$$

- Tendència: és la component general de la sèrie i pot considerar-se com el moviment global de la sèrie a llarg termini. Si va cap amunt o cap avall.

- Estacionarietat: mesura la presència de cicles, de pujades/baixades

    - Variacions cícliques: oscil·lacions periòdiques que es produeixen amb una freqüència superior a un any, on es combinen etapes de prosperitat amb etapes de depressió.

    - Variacions estacionals: fluctuacions de periodicitat inferior a l'any i reconeixible tots els anys, solen estar relacionades amb la climatologia o el comportament dels agents econòmics al variar l'època de l'any.

- Aleatorietat: mesura desviaments respecte la tendència i a l'estacionarietat. Són les variacions erràtiques, irregulars o residuals i recullen la variabilitat en el comportament de la sèrie deguda a petites causes impredictibles.

A principis de 2017, Facebook va publicar la llibreria 'Prophet' que permet realitzar prediccions de dades a partir d'una sèrie temporal. Es recomana utilitzar aquesta llibreria si es disposen de dades diàries. Tot i que, el millor d'aquest llibreria és que no requereix massa coneixements previs de predicció de dades, ja que troba automàticament tendències estacionals sota les dades i ofereix un conjunt de paràmetres fàcils d'entendre. Altres mètodes com Forecast o Arima requereixen més coneixements i habilitats de configuració que no són l'objectiu de la pràctica. Per això, emprarem Prophet per a la predicció de les dades.

a. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Analitzarem els conjunts de dades següents:

- taula amb el total de vendes per dia (date, total_vendes) per analitzar la serie temporal. Cal normalitzar les dades i aplicar l'algorisme prophet per analitzar la tendència, estacionalitat i aleatorietat. Amb aquest anàlisi volem veure quina tendència té la serie temporal.

- a partir del conjunt de dades de prova (test.csv), per cada registre amb el identificador de la tupla, botiga (shop_id) i producte (item_id), crearem:
  - una taula amb el total de vendes per dia, per cada shop_id i item_id concret
  - predicció de les vendes del més de novembre 2015, d'aquest shop_id i item_id concret
  - emmagatzemarem el resultat en una nova taula amb el identificador de la tupla (ID) i la predicció de venda d'aquest shop_id i item_id concret
  - aquesta taula resultant és la que guardarem en un nou fitxer:

b. Comprovació de la normalitat i homogeneïtat de la variància.

Una de les maneres més fàcils d'examinar la normalitat dels residus és mitjançant l'aplicació de l'estadístic Jarque-Bera.

c. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Per validar la serie temporal podem emprar un contrast d'hipòtesi (prova paramètrica)

- H0 (hipotesi nul·la): no hi ha tendència en aquesta serie temporal.
- H1: hi ha tendència en aquesta serie temporal.

https://help.xlstat.com/customer/es/portal/articles/2062453-what-is-a-statistical-test-?b_id=9283

Amb els components del prophet podem demostrar l'existència de tendencia en les dades:
https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet

Podem fer prova no paramètrica de tendències Mann-Kendall. No és tan robusta com la prova paramètrica de contrast d'hipotesi.
https://help.xlstat.com/customer/es/portal/articles/2062303

Altres tests que es poden aplicar són: Encopassing test
https://msperlin.github.io/2017-03-05-Prophet-and_stock-market/

5. Representació dels resultats a partir de taules i gràfiques.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

Recursos web:
https://github.com/kazimanil/predict-future-sales

https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet

https://www.kaggle.com/c/competitive-data-science-predict-future-sales/discussion/54949

https://techtravelo.wordpress.com/2014/03/21/installing-package-forecast-in-r/

https://estadisticaorquestainstrumento.wordpress.com/2013/04/30/tema-25-analisis-de-series-temporales/

http://www.doctormetrics.com/2017/04/27/introduccion-al-forecasting-con-r-statistics/#.WvibDRyxU5k

https://www.kdnuggets.com/2018/03/time-series-dummies-3-step-process.html

https://www.kaggle.com/vyordanov/simple-prediction-approach-to-get-you-in-top-70/code

https://www.otexts.org/fpp/9/4

https://www.kaggle.com/jagangupta/time-series-basics-exploring-traditional-ts

https://rstudio-pubs-static.s3.amazonaws.com/289564_7557e57a8aac42b1a8ca434689ee3cff.html

Exemples prophet:
https://cran.r-project.org/web/packages/prophet/prophet.pdf

https://msperlin.github.io/2017-03-05-Prophet-and_stock-market/

https://www.kaggle.com/vyordanov/simple-prediction-approach-to-get-you-in-top-70

https://www.kaggle.com/kazimanil/predicting-future-sales
https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials

Exemples timekit:
http://www.business-science.io/code-tools/2017/05/02/timekit-0-2-0.html

Normalitat residus:
http://www.eumed.net/ce/2011a/chai.htm
https://www.quality-control-plan.com/StatGuide/n-dist_alts.htm

Criteris de valoració

Tots els apartat són obligatoris. La ponderació dels exercicis és la següent:

- Els apartats 1, 2 i 6 valen 0,5 punts.
- Els apartats 3,5 i 7 valen 2 punts.
- L'apartat 4 val 2,5 punts.

Es valorarà la idoneïtat de les respostes, que han de ser clares i completes. Les diferents etapes han d'estar ben justificades i acompanyades del codi corresponent. També es valorarà la síntesi i claredat, a través de l'ús de comentaris, del codi resultant, així com la qualitat de les dades finals analitzades.