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 pretèn 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 diaries de l'empresa **1C Company**, una de les companyies de programari russes més grans.

L'objectiu és predir la quantitat total de productes que es venen a cada botiga per al conjunt de proves per tal de:
- 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 proveïr 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.

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

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

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

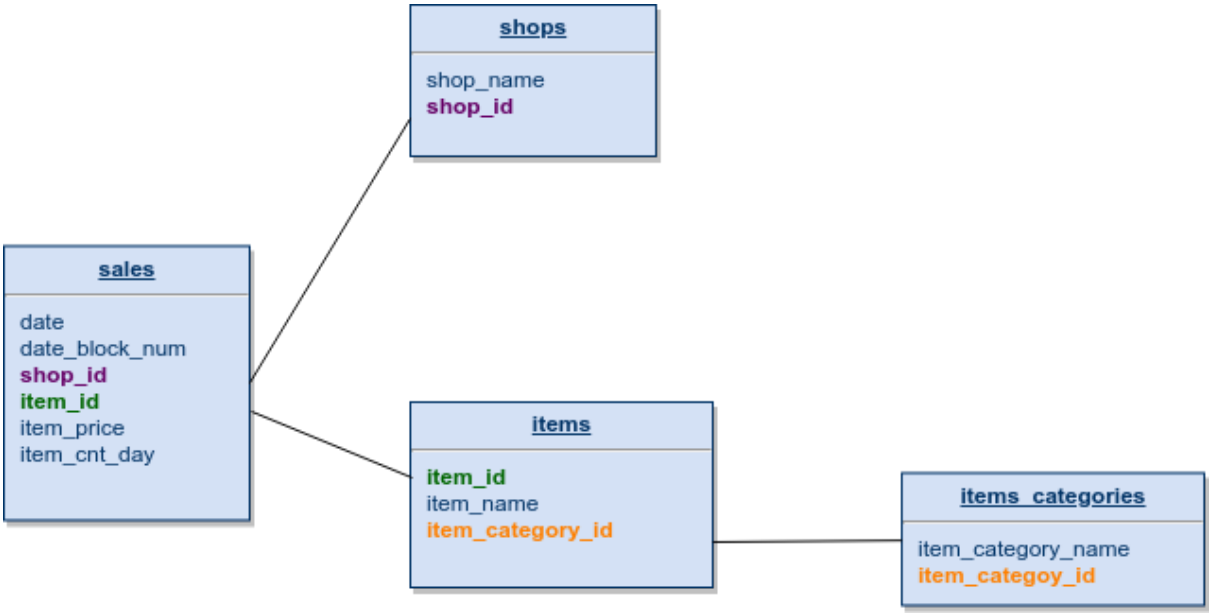
Descripció dels fitxers

- ***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'inicialitzar amb el canvi d'any. Per exempl, gener de 2013 és 0, febrer de 2013 és 1, ..., l'octubre de 2015 és 33.

Model E-R de les vendes per botiga



Amb l'objectiu de predir la quantitat total de productes que es venen a cada botiga, en primer lloc construirem un fitxer que integri les dades referents a productes, botigues i vendes.

Per a fer l'estudi de predicció de vendes, podem ometre els noms dels productes, així com de les botigues i de les categories.
Tampoc ens fa falta saber el preu dels productes.
Per que fa a la data la convertirem en 2 camps -> any i mes i ometrem el camp date_block_num
De manera que només amb els identificadors del fitxer sales, la data i la quantitat venuda ja en tindrem prou.

3. Neteja de les dades.
    a. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
    b. Identificació i tractament de valors extrems.

4. Anàlisi de les dades.
    a. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
    b. Comprovació de la normalitat i homogeneïtat de la variància.

   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.

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.

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.