Tipologia i cicle de vida de les dades

Pràctica 2

Mireia Calzada i Noemi Lorente

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"](https://www.coursera.org/learn/competitive-data-science/home/welcome), 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**](http://1c.ru/eng/title.htm), 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 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.

Es pot consultar el dataset a <https://github.com/kazimanil/predict-future-sales>

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

## **Descripció dels fitxers**

***items.csv***: informació addicional sobre els articles / productes.

***item\_categories.csv***: informació complementària sobre les categories d'elements.

***shops.csv***- informació complementària sobre les botigues.

***sales\_train.csv****: Dades històriques diàries de gener de 2013 a octubre de 2015.*

***test.csv***: Conjunt de proves. Cal preveure les vendes d'aquestes botigues i productes per a novembre de 2015.

***sample\_submission.csv***: un fitxer de presentació de mostres en el format correcte.

## **Camps de dades**

**ID**: una identificació que representa una tupla (de botiga, d'element) dins del conjunt de proves

**shop\_id** - identificador únic d'una botiga

**item\_id** - identificador únic d'un producte

**item\_category\_id** - identificador únic de categoria d'element

**item\_cnt\_day** - nombre de productes venuts. Esteu predient una quantitat mensual d'aquesta mesura

**item\_price** - preu actual d'un element

**data** - data en format dd / mm / aaaa

**data\_block\_num** - un número de mes consecutiu, usat per a la comoditat. El gener de 2013 és 0, febrer de 2013 és 1, ..., l'octubre de 2015 és 33

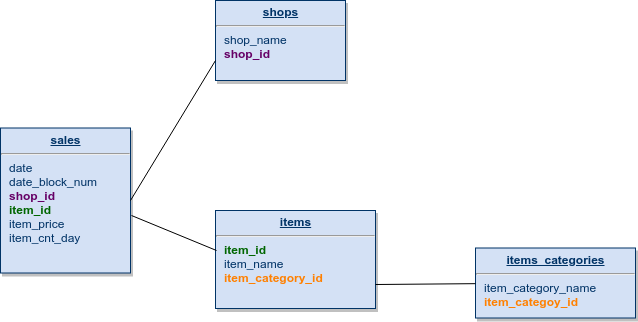
**item\_name** - nom de l'element

**shop\_name** - nom de la botiga

**item\_category\_name** - nom de la categoria de l'element

1. Integració i selecció de les dades d’interès a analitzar.

## 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.

1. Neteja de les dades.
   1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
   2. Identificació i tractament de valors extrems.
2. Anàlisi de les dades.
   1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
   2. Comprovació de la normalitat i homogeneïtat de la variància.
   3. 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.
3. Representació dels resultats a partir de taules i gràfiques.
4. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
5. 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.

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.