10-6-2019

TiCD- PRÀCTICA 2

Descripció : Neteja i validació de dades

Pràctica 2 – Neteja I validació de dades Carles Maggi -Joan Maggi

UOC-TCD- PrÀctica 2

Contenido

[1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre? 3](#_Toc11165412)

[2. Integració i selecció de les dades d’interès a analitzar. 4](#_Toc11165413)

[3. Neteja de les dades. 4](#_Toc11165414)

[3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos? 4](#_Toc11165415)

[3.2. Identificació i tractament de valors extrems. 5](#_Toc11165416)

[4. Anàlisi de les dades. 5](#_Toc11165417)

[4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar). 5](#_Toc11165418)

[4.2. Comprovació de la normalitat i homogeneïtat de la variància. 6](#_Toc11165419)

[4.3. 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. Aplicar almenys tres mètodes d’anàlisi diferents. 6](#_Toc11165420)

[5. Representació dels resultats a partir de taules i gràfiques. 6](#_Toc11165421)

[6. Resolució del problema. A partir dels resultats obtinguts, quines son les conclusions? Els resultats permeten respondre al problema? 6](#_Toc11165422)

[Primer problema 6](#_Toc11165423)

[Segon problema 7](#_Toc11165424)

[7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s’ha realitzat la neteja, anàlisi i representació de les dades. 8](#_Toc11165425)

[TAULA DE CONTRIBUCIONS 8](#_Toc11165426)

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

Podem trobar els datasets en el següent repositori :

<https://github.com/joanmaggigo/practica2/CSV>

La descripció dels fitxers inicials és la següent :

**HotelsBarcelonaBooking :**

IdHotel : BigInt que identifica de manera unívoca el hotel

Estrelles : Int Número d'estrelles, en cas que n'hi hagi

Nota : Float mitja de l'hotel

Nom : String Nom de l'hotel

Link : String Url de la pàgina de l'hotel

**ComentarisXHotelsBarcelonaBooking :**

IdHotel : BigInt Identifica de manera unívoca el hotel

IdIteració : Int Primera part que identifica un comentari (iteració llista comentaris)

Index : In\* Segon part d'identificació de comentari (número de comentari dins la iteració)

NomHotel : String Nom de l'hotel

Nota : Float Nota que en qualifica el comentari de l'hotel

Comentari Postiu : Text Comentari positiu si n'hi ha

Comentari Negatiu : Text Comentari negatiu si n'hi ha

Data Comentari : Date Data enregistrada del comentari PRAC 1 -Tipologia i Cicle de Dades

**CategoriesXComentariBooking :**

IdHotel : BigInt Identifica de manera unívoca el hotel

IdIteració : Int Primera part que identifica un comentari (iteració llista comentaris)

Index : Int Segon part d'identificació de comentari (número de comentari dins la iteració)

Categoria : Text Categoria amb la que s'ha categoritzat el comentari.

**Importància :** Ens cal perquè és el recopilatori de les puntuacions donades pels usuaris als diferents hotels, que serà la base del nostre treball

**Pregunta 1:**

Perceben el usuaris alguna una diferencia de qualitat de servei entre estiu i hivern?

**Pregunta 2:**

Podríem recomanar un hotel a partir de les puntuacions dels usuaris per una combinació concreta de paràmetres relacionades, amb tipus de viatge, acompanyament, tipus d’habitacions, etc..?

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

Un cop definit l’objectiu , caldrà integrà totes les dades en un sol Data set per poder fer totes les probes i testos necessaris. Quedarà un data set de 16081 observacions i 14 variables.

El data set de treball tindrà les següents atributs:

**nomHotel** : String, Nom del Hotel

**notapersona** : num ,Nota Posada al hotel per la Persona

**data** : Date, Data del comentari

**estrelles** : num, Estrelles del Hotel

**notaHotel** : Nota mitja concedida al Hotel per Booking

**TipusHabitacio**  : String, Tipus d’Habitació

**DuradaEstancia**  : String, Rang de la estància

**Nits** : num, Número de nits d’estància

**ProcedenciaComentari** : String, Lloc de des de on s’ha fet el comentari

**TipusViatge**  : String, Tipus de Viatge

**Acompanyament** : String, Tipus d’acompanyament

**ViajaConMascota** : String, Si viatja en mascota

**Season**  : String, Temporada de l’estada

**Recomanar :** num **,** Si es recomana o no l’Hotel

# 3. Neteja de les dades.

## 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En el moment de treballar amb el data set d’hotels, veiem que els que el nombre d’estrelles tenen un 0 es refereixen a apartaments. Com l’objectiu de la pràctica és analitzar les variacions del servei entre estiu i hivern i veure si podem recomanar un hotel a les persones en funció de X regressors, considerem oportú eliminar tots els comentaris dels hotels que són apartaments, és a dir estrelles = 0

La variable viajaconMascota, té NA’s. Molts. El que entenem es que hi ha una definició quan viatja amb mascota, però no hi ha definició quan no porta mascota. Per tant entenem la categoria com a binaria i substituïm els valors amb S/N en funció de si tenim valor o no.

La variable Tipusdeviatge representa un 10% del volum total, per tant considerem que podem tractar-ho com una descripció de categoria nova, ja que té un volum prou important.

La variable ProcedenciaComentari té un volum elevat. Entenent la variable, que en els casos que ve informada ve amb el valor “EnviadoporMóbil” entenem que només hi ha dues maneres d’enviar els comentaris, por mòbil o por web.

La variable Nits i el Tipus D’habitació (Duradaestancia és homòleg a Nits, però categoritzat) veiem que són els mateixos elements de la mostra. Entenent el significat de les variables, considerem que són errors, és a dir, que no han informat ni del tipus d’habitació ni el nombre de nits a l’hotel, per tant, considerem que cal eliminar-ho.

## 3.2. Identificació i tractament de valors extrems.

El dataset té tres variables numèriques, la primera és notahotel (que és la nota de booking) i es repeteix per cada un dels comentaris que ha rebut el hotel. Seria un error gestionar els outliers dins del data set construït (ja que el pes de la nota anirà en relació al volum de comentaris) tanmateix, no la considerem per fer els nostres anàlisis posteriors, només com una variable que ens ajuda a definir una nova variable inventada (recomana la persona l’hotel o no, si la nota que posa es superior a la mitjana de l’hotel). En qualsevol cas no té sentit eliminar outliers en aquesta variable.

Pel que fa a la notapersona, és la nota que posa la persona a un hotel. No té sentir mirar-ho pel conjunt del data set, sinó utilitzem el nom de l’hotel com a filtre per analitzar si hi ha o no outliers.

Mirarem els valors extrems de la puntuació donada per un client en referencia a un mateix hotel, no pas tenint en compte totes les puntuacions.

Eliminarem els registres que sobrepassen 3 desviacions estàndards la nota donada per les persones. Ho fem a nivell d'hotel, i no pas a nivell general, perquè entenem que si un hotel és molt dolent potser tothom el puntua amb 1 o 2, que de segur serien outliers si els consideréssim en la totalitat de les puntuacions de tots els hotels, però serien valors força normals a la puntuació pròpia del hotel.

Nombre de nits, veiem que en termes generals hi ha 242 outliers (per sobre de 1,5\*IQR+QuartQuantil). Fent una inspecció visual, els veiem vinculat a un nombre determinat d’hotels i ens sembla possible les dades que es manifesten. Per tant els considerem com a valors raonables que no cal el tractament i reforça el fet d’haver creat la categòrica DuracióEstància. El que considerem es que el nombre de nits de per sí potser no és un variable interessant amb la que tractar.

# 4. Anàlisi de les dades.

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

1) Primer hem fet un lectura dels 3 datasets i hem necessitat fer la fusió dels mateixos.

2) Hem vist que la notahotel no és analitzable en el resultat del dataset de treball, perquè és una nota mitjana que apareix tantes vegades com comentaris hi ha a l’hotel , però la mantenim pels nostres plans d’intentar fer un aconsellador d’Hoteles

2.1) També veiem que la variable Nits, és més interessant tractar-la com en una categorització que com un numèric.

3) Creem una primera hipòtesis i és que el servei dels hotels és igual a l’estiu que a l’hivern. Per fer això, necessitem crear una nova variable que ens determini el trimestre de l’any. Entenem que tot i que la data és la data en la que es fa el comentari, aquesta guarda una relació de proximitat amb la data de la estància.

3.1 ) Analitzem la notapersona en termes generals. I fem el test en terme general

3.2) Particularitzem per cada categòrica, entenent-la com un filtre

3.3) Treballem les conclusions

4) Definim un model de predicció per fer un motor de recomanació.

4.1 ) Provem de fer una predicció de la nota en funció de les categòriques

4.2) Si ens surt que no és plausible

4.2.1) Creem una nova variable, recomana (S/N) si la nota que posa la persona es superior a la nota de l’hotel.

4.2.2) Creem un model per cadascun dels hotels.

4.3) Conclusions

## 4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Es pot comprovar en l’exercici fet en R

## 4.3. 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. Aplicar almenys tres mètodes d’anàlisi diferents.

Es pot comprovar en l’exercici fet en R

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

Es pot comprovar en l’exercici fet en R

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

## Primer problema

El primer problema que volíem estudiar era si hi havia diferències significatives entre la nota d’estiu i d’hivern, en termes generals i després filtrant per cadascuna de les categòriques descriptives.

Hem trobat que, la nota que posa la persona no segueix una distribució normal (observable tant gràficament com a través de les proves de Shapiro). Tot i així hem comprovat l’homoscedasticitat, que tampoc es dona.

Sense tenir en compte les categòriques**, hi ha un diferència de nota significativa entres Hivern i Estiu**, i si mirem les mitjanes, aquesta és menor a l’estiu que a l’hivern**, pel que podem concloure que hi ha pitjor servei, o és percebut pitjor per el client, a l’estiu que a l’hivern**.

Si fem filtres específics per cada categòrica, i apliquem test de normalitat per determinar quin test a aplicar per veure la diferència entre mitjanes, veiem els següents resultats

#### SEGMENTACIÓ PER ESTRELLES

Existeixen diferències significatives entre les notes pels hotels d’1 estrella a 4 estrelles. No és així pels hotels de 5 estrelles. Això ens porta a pensar que en els hotels de 5 estrelles el servei és igual de bo a l’estiu que a l’hivern, per tant, es probable, si tinguéssim la dada, que existís una forta relació entre preu i qualitat del servei.

Observem que si el nostre llindar fos 0.01 en comptes de 0.05 els hotels d’1 estrella tampoc tindrien una diferència significativa entre estiu i hivern.

#### SEGMENTACIÓ PER TIPUS HABITACIÓ

En aquest cas sembla que la gent que viatge amb una habitació doble o quàdruple, si que nota una diferència en el servei d’estiu a hivern. Altres és al llindar.

#### SEGMENTACIÓ PER DURADA DE VIATGE

En aquest cas veiem que com més llarga és la estància menys diferència es percep en el servei entre el servei de l’estiu i de l’hivern, a partir de 6 nits no es percep diferència entre la nota d’estiu i la d’hivern

#### SEGMENTACIÓ PER ACOMPANYAMENT

Sembla que només aquells que viatgem amb un grup d’amics no percebin una diferència entre el servei de l’hotel a l’estiu i hivern. Potser aquesta categoria ens indica que és un grup de gent que potser viatgen no pas pels serveis de l’hotel....

#### SEGMENTACIÓ PER TIPUS DE VIATGE

En aquest cas tots els grups perceben diferència significativa entre el servei d’estiu i el servei que es dona a l’hivern.

És a dir, que veiem que totes les categòriques, a excepció de tipus de viatge, i dins de les categòriques uns grups en concret, perceben canvis en el servei d’estiu a hivern.

## Segon problema

El segon problema que hem plantejat és si seriem capaços de recomanar un hotel segons la descripció de les categòriques. Per això, hem pensat que fora bo provar de fer una regressió que ens ajudés a modelar la predicció de la nota en funció de les categòriques i una variable continua.

Hem triat les següents categòriques: TipusHabitació, TipusViathe, Acompanyament i Temporada i la continua Nits.

Al mirar els resultats en general hem vist que teníem una explicació de la variabilitat R^2 molt baixa, pel que ja hem vist que ens faltaven més regressors per poder explicar la variabilitat de la nota. A partir d’aquí hem decidit canviar d’estratègia.

Hem creat una nova variable, recomana, que si la nota que ha posat l’usuari és major que la nota mitjana aleshores recomana l’hotel, si és menor no el recomana i hem realitzat un enfoc de predicció amb una regressió logística.

Primer hem mirat en termes generals com expliquen els regressors aquesta nova variable, hem vist que hi ha certes relacions explicades en el document. Tot i així, aquest exercici respon més a un objecte de la pràctica que la decisió que prendríem a l’inici que és realitzar una regressió per cadascun dels hotels, és a dir, segmentant per hotel obtenir un model per cada hotel.

Un cop obtinguda aquesta regressió, hem vist que en termes generals les regressions ajustades a cada hotel expliquen millor el model de recomanar o no aquell hotel. No és perfecte, però tenim un resultat que hem tirat endavant perquè volíem tancar una primera versió del que podria ser un producte de recomanació.

Hem de tenir en compte que:

1. Hem eliminat aquells hotels que no tenien com a mínim representacions tant de recomanacions com de no recomanacions
2. Hem obviat el fet que en el moment de fer segmentacions potser no existeix una mostra prou representativa suficient que garanteixi que existeix com a mínim una observació de les descripcions possibles de les categòriques. Això és resoldria ampliant el conjunt de dades.
3. Tampoc hem entrat en el procés de valorar si en el conjunt d’observacions de cada hotel tenim una mostra balancejada adequadament de casos en positiu i casos en negatiu.

Dit això, un cop obtingut un model per cada hotel i el valor òptim que ens maximitza la corba ROC (és a dir cerca l’equilibri més bo entre sensibilitat i especificitat) determinant a partir de quin valor acceptar el cas com a positiu. Hem procedit a definir l’algoritme de recomanació.

L’algoritme de recomanació a partir d’una entrada d’un sol registre amb els valors que esperats (TipusHabitacio, Nits, TipusViatge, Acompanyament i Season) executa la predicció amb el model respectiu per cada hotel, calcula la diferència entre el valor òptim i el valor predit, i si és positiu, endreça de major a menor per la diferència i la nota mitja, retornant dues llistes, una amb els hotels recomanats i l’altra amb els hotels no recomanats.

# 7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s’ha realitzat la neteja, anàlisi i representació de les dades.

Es pot trobar el codi en

<https://github.com/joanmaggigo/practica2>

# TAULA DE CONTRIBUCIONS

|  |  |
| --- | --- |
| **Contribucions** | **Firma** |
| **Investigació Prèvia** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |
| **Redacció de les Respostes** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |
| **Desenvolupament del Codi** | CM,JM (CM->Carles Maggi, JM->Joan Maggi) |