**Resolució de la Pràctica 2**

En aquesta segona pràctica es presenta una anàlisis sobre el data set de “Titanic”. Aquest data set conté informació sobre els passatges del RMS Titanic, el famós transatlàntic britànic que es va enfonsar a l’abril del 1912 durant el seu primer viatge des de Southampton fins a Nova York. En concret, hi trobem informació sobre l'edat, el gendre i les característiques socio-econòmiques dels passatgers, així com també si van sobreviure al naufragi o no.

Aquest data set es troba disponible amb llicència pública en el següent enllaç: <https://www.kaggle.com/c/titanic>

**Descripció del dataset**

**Detalls sobre el data set**

De forma més detallada, el data set es troba disponible en format “*comma separated values (csv)*” i conté les següents variables:

* PassengerId: es tracta d’una variable qualitativa ordinal, on cada valor representa el nom d’un passatge de manera numèrica.
* Survived: la segona variable indica si el passatger va sobreviure o no al naufragi i ho representa amb una variable qualitativa binària, on 0 indica que no ho va fer, mentre que 1 sí va sobreviure.
* Pclass: variable qualitativa ordinal, que pot prendre qualsevol valor enter entre 1 i 3. Cada valor indica el tipus de classe en la que viatjava el passatger.
* Name: variable qualitativa nominal que indica el nom del passatger.
* Sex: variable qualitativa binaria, que indica el gendre del passatger.
* Age: indica l’edat del passatger, pel que es tracta d’una variable quantitativa discreta.
* SibSp: indica el número de germans, germanes i/o germanastres de cada passatger en el vaixell. Es tracta d’una variable quantitativa discreta.
* Parch: Nombre de pares o fills en el vaixell, pel que també és una variable quantitativa discreta.
* Ticket: Número identificador del bitllet del passatger, pel que és una variable qualitativa ordinal.
* Fare**:** Preu pagat pel bitllet. Per tant, es tracta d’una variable quantitativa contínua.
* Cabin: Identificador de la cabina assignada a cada passatger. Es tracta d’una variable qualitativa ordinal.
* Embarked: Port en el que va embarcar el passatger, pel que és una variable qualitativa nominal.

El data set que utilitzarem serà l’anomenat “train.csv”, el qual conté 891 variables. El subdividrem en dos data sets, un per entrenar les dades amb el(s) model(s) escollit(s) per tal de que aquest aprengui i després poder-lo utilitzar per realitzar prediccions. El segon sub-data set s’utilitzarà per estudiar la bondat del model, mitjançant la comparació del valor predit pel model i el valor real de supervivència.

Nota: No és possible utilitzar el data set “test.csv” disponible a l’enllaç mencionat anteriorment, ja que no disposa dels valors reals de supervivència. Per tant, no seria possible analitzar la bondat de predicció del model entrenat. Aquest l’utilitzarem només per a fer prediccions.

**Preguntes plantejades per a la resolució de l’activitat**

* Quin és el grup de passatgers que va sobreviure més en percentatge del total?
* És a dir, quines variables ajuden a predir de manera més precisa la supervivència d’un passatger?
* Hi ha alguna correlació entre l’edat, el gendre, el tipus de classe i la grandària de la família?
* Els passatgers més joves tenien una probabilitat de supervivència més alta que la resta?

**Integració i selecció de les dades d’interès a analitzar**

El data set que emprarem principalment per dur a terme la anàlisis de les dades és el anomenat “train.csv”. Amb la totalitat del data set, farem una anàlisi descriptiva de les dades.

Per altra banda, el 70% dels s’utilitzarà per entrenar el model i, el 30%, per testejar la seva bondat.

Per altre costat, el data set “test.csv” s’utilitzarà només per a realitzar prediccions.

**Neteja de les dades**

**Objectiu**

En primer lloc, crearem dues variables noves, les quals generarem a partir d’atributs que venen per defecte.

Primerament, crearem una anomenada “FamilySize”, la qual indicarà la grandària de la família, on el valor mínim serà 1, en el cas de que el passatger viatgés sol. Per tal realitzar-la, utilitzarem les variables “SibSp” i “Parches”, les quals indiquen el número de germans i de pares que viatjaven amb el passatger, respectivament. Amb aquesta nova variable volem veure si la mida de la família podria tenir un impacte en la supervivència del passatger, així ajudant a respondre una de les preguntes plantejades.

Per altre costat, crearem una altra variable, anomenada “Title”, que indicarà el títol del passatger o tripulant. Aquesta variable la crearem a partir de l’extracció de la primera part del nom (“Name”) del passatger.

En segon lloc, analitzarem si alguna de les variables que venen per defecte o de les creades per l’usuari contenen valors no contestats o buits. En el cas de que sí, els valors buits o no contestats s’han de reemplaçar segons algun criteri matemàtic o bé, eliminar la totalitat de l’atribut, en cas de que sigui convenient. Aquest cas es podria donar, per exemple, quan la majoria de les dades d’aquell atribut es troba sense contestar.

A continuació, haurem d’analitzar i decidir quines variables considerem rellevants per a la anàlisi i quines no, les quals podrem prescindir.

I, finalment, analitzarem i tractarem els valors extrems.

**Resolució**

Valors buits o no contestats

Un cop creades les noves variables FamilySize i Title, busquem si alguna de totes les variables contenen algun valor buit o no contestat. De fet, tres variables en contenen algun. La variable Age conté 177 valors no contestats, els quals representen el 19.86% del total; la variable Embarked conté 2 valors buits, que representen el 0.22% del total i, finalment, la variable Cabin conté 687 valors buits, que representen el 77.1% dels casos.

Pel que fa a la variable Embarked, substituïm els valors buits pel valor més comú. En canvi, per la variable edat, procedim d’una altra manera, ja que si ho substituíssim pel valor més comú o bé per la mitjana, podria fàcilment portar a anàlisis esbiaixats, ja que un nombre bastant elevant (quasi el 20%) dels valors es troben no contestats.

En aquest cas, substituirem els valors NA en funció dels valors semblants del passatger, segons la classe social i el seu gendre, ja que aquestes dues variables no es comporten d’igual manera.

Finalment, prescindirem de la variable Cabin, ja que la majoria dels seus valors es troben sense contestar.

Reducció de la dimensionalitat

Primerament, prescindirem dels atributs PassengerId, Name i Ticket, ja que no aporten cap valor afegit en la anàlisis, ja que es tracta de variables úniques i, per tant, irrellevants.

A més a més, les variables Title, SibSp i Parche s’han emprat per a altres variables. Per tant, indirectament ja les estem considerant en les noves variables creades (FamilySize en el cas de SibSp i Parche; i l’estimació de l’edat de valors NA pel Title).

Finalment, podem també prescindir de la variable Fare. Això és degut a que les dones de primer classe van pagar un bitllet més car que la resta. Una de les raons és que hi ha una gran probabilitat que aquest grup va haver d’assumir els bitllets dels infants. Ja que la raó no és clara, decidim prescindir també d’aquesta variable.

Valors extrems o outliers

Ja que la variable Age és la única variable numèrica, només en aquest cas se’n poden donar. A través d’un gràfic de caixes (box plot), observem que els outliers es donen únicament en el sector de la gent gran. Dins dels valors extrems, el més elevat és el 80 i, el més inferior és el 61.

Donat que es tracta d’un interval d’edats totalment raonable, considerarem tots els valors dins de la variable Age com a vàlids.

**Anàlisi de les dades**

**Objectius**

En aquest apartat, l’objectiu serà analitzar les relacions de dependència entre les diferents variables i crear diferents models de regressió, així comparant-los per trobar-ne l’òptim.

Primerament, però, caldrà analitzar la distribució de les variables, així com la heterogeneïtat de les variables. D’aquesta manera, podrem decidir quins són els models que més s’ajusten a la nostre anàlisis.

**Resolució**

Normalitat

Comprovem només la distribució de la variable Age, ja que és la única variable numèrica que tenim. Per tal de dur-ho a terme, hem representat l’atribut en un qqplot. Malauradament, el gràfic no és prou precís, ja que els valors del centre podrien portar a la conclusió que la variable segueix una distribució normal. En canvi, els extrems indiquen el contrari. A més a més, pel teorema central del límit, podríem assumir normalitat. Ja que no és del tot clar, hem realitzat també un contrast d’hipòtesi a través del Shapiro-Wilk test. Ja que la hipòtesi nul·la es rebutja amb un nivell de significació del 5%, afirmem que la variable no segueix una distribució normal.

FALTA FAMILYSIZE 🡪 TAMBÉ ÉS NUMÈRICA

Homogeneïtat

En el següent pas, hem comprovat si la variància de l’atribut Age és significativament diferent que la de la variable Survived. Amb un nivell de significació del 5%, rebutgem la hipòtesis nul·la, pel que podem concloure que les dues variables tenen variàncies significativament diferents. A més a més, l’Interval de Confiança es troba per sota de 1.

Relacions de dependència

En aquest apartat, hem analitzat diferents relacions de dependència entre la variable Survived i totes les altres variables, sempre de forma individual. Per dur-ho a terme, hem emprat un Chi-Squared test, on la hipòtesis nul·la assumeix independència entre la variable Survived i cada variable.

En tots els casos, la hipòtesis nul·la es rebutja i s’accepta la alternativa. Per tant, la variable Survived té una relació de dependència o associació amb un nivell de significació del 5% amb cadascuna de les variables que estem utilitzant per a realitzar la anàlisis.

Contrast d’hipòtesis

Realitzem un contrast d’hipòtesis, per veure si els passatgers amb una edat inferior a 20 anys sobreviurien o no al naufragi.

El contrast d’hipòtesis l’hem realitzat a través del test U de Mann.Whitney, assumint que les mostres són independents. No podem utilitzar el test t-student, ja que necessitaríem que les dades es distribuïssin normalment i, a més a més, les variàncies de les variables haurien de mantenir-se constants en el rang observat d’alguna altra variable. En aquest test, la hipòtesis nul·la assumeix que els passatgers amb una edat inferior a 20 anys no tenen més probabilitat de supervivència.

Un cop realitzat el test, veiem que la hipòtesis nul·la es rebutja amb un nivell de significació del 5%. Per tant, les probabilitats de supervivència d’una persona amb una edat inferior a 20 anys és més elevada que la resta.

Models de regressió

Abans de crear qualsevol model, hem dividit el data set en dos, un amb 70% dels casos i, l’altre, amb la resta. El primer l’hem utilitzat per a crear diferents models explicats a continuació, mentre que l’altre per testejar la bondat dels models.

En aquest punt, hem creat diversos models de regressió logística. Primerament, hem realitzat un model per a cada variable. És a dir, un model de regressió amb la variable dependent Survived i cada atribut com a variable independent. Per tant, hem realitzat en total quatre models, on es tracta de trobar en quin percentatge es modifica la probabilitat de supervivència en funció de cada atribut.

Pel que fa al gendre, el fet que els homes sobrevisquin és un 93% més baix que les dones. Pel que fa al port d’embarcació, els passatgers que ho van fer pel port Q tenen una probabilitat de supervivència més baixa per un 53.5% respecte els del port C. Per altre costat, els del port S tenen una probabilitat més baixa que els del C per un 63.3%. Pel que respecte a la grandària de la família, els passatgers solitaris tenen un XX% de supervivència menys que els de les famílies petites i, els passatgers de famílies grans tenen un XX% de supervivència més baix que el de les famílies més petites. Finalment, els passatgers joves tenen una probabilitat d’un 73.3% inferior de supervivència respecte els infants; els passatgers adults d’un 58.3%; els adults avançats d’un 67.8% i, la gent gran, d’un 93.9%.

En segon lloc, hem dut a terme un model de regressió logística multivariable. És a dir, hem realitzat un únic model, on la variable dependent era la supervivència, mentre que les variables independents han estat la resta.

En aquest model, hem pogut observar que el valor AIC és inferior a 500, pel que la precisió del model no era gaire alta. A més a més, hem pogut observar que la variable Embarked no és significativa, és a dir, no ajuda a la predicció de la supervivència d’un passatger. Per tant, hem realitzat un segon model de regressió logística multivariable, on hem exclòs l’atribut Embarked.

En aquest segon model, el valor AIC ha millorat lleugerament, però el resultat no és del tot satisfactori. A més a més, hem observat que la variable FamilySize no és significativa en cap nivell. Per tant, hem optat per realitzar un tercer model, excloent també l’atribut FamilySize.

En aquest últim model, el valor AIC millora per sobre dels 570 on totes les variables són significatives, al menys a un nivell del 10%. Per tant, dels tres models utilitzats, podem concloure que aquest és el més precís de tots.

Finalment, hem comprovat la bondat del model, així testejant la predicció a partir del 30% de les dades del data set. Per tal de veure-ho, hem dut a terme una matriu de confusió, la qual ens indica que el model té un 77% de precisió. És a dir, en el 77.6% dels casos, el model predirà de forma correcta si el passatger, donades les variables, sobreviuria o no al naufragi.

**Representació del resultat**

**Objectius**

**Resolució**

**Resolució del problema**

**Participació dels integrants**

|  |  |
| --- | --- |
| Contribucions | Firma dels integrants |
| Investigació prèvia | Jordi Dil i Giró / Carlota Font Castell |
| Redacció de les respostes | Jordi Dil i Giró / Carlota Font Castell |
| Desenvolupament del codi | Jordi Dil i Giró / Carlota Font Castell |

**Bibliografia**

1. Titanic: Machine Learning from Disaster | Kaggle [Internet]. Kaggle.com. 2020 [citat 7 de desembre 2020]. Disponible en l’enllaç: <https://www.kaggle.com/c/titanic/overview>