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.

En primer lloc, ajuntarem els dos data sets disponibles per tal de fer un tractament i neteja de les dades. A continuació, utilitzarem les dades del data set “train.csv”, ja que conté les dades de supervivència, i el subdividrem en dos data sets, un per entrenar les dades amb el(s) model(s) escollit(s) per tal de que aquest aprengui. 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.

Finalment, emprarem les dades del data set “test.csv” per a realitzar prediccions amb el model més precís.

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 la combinació dels dos data sets disponibles, "train.csv" i "test.csv". Amb la totalitat del data set, farem una anàlisi descriptiva de les dades. En aquet data set total hi afegirem una variable extra anomenada que ens permetrà distingir ràpidament si les observacions es tracten del fitxer original train o test.

Per altra banda, del data set "train.csv", el 75% de les dades s'utilitzarà per entrenar el model i, el 25%, 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, comprovarem si existeixen valors duplicats i si hi ha valors sense contestar o en blanc. En segon lloc, observarem la distribució de cada variable en funció de Survived.

A continuació, analitzarem si és convenient crear noves variables a partir dels atributs que venen per defecte. crearem dues variables noves, les quals generarem a partir d'atributs que venen per defecte.

Després, analitzarem i tractarem els valors extrems.

A continuació, en el cas que existeixin valors buits o no contestats, analitzarem quina és la millor opció per a tractar-los. Per exemple, els podrem reemplaçar per algun criteri matemàtic o, bé, eliminar la totalitat de l'atribut.

I, finalment, analitzarem la possibilitat de discretitzar algunes variables, en el cas de ser convenient.

Resolució

<u>Valors buits o no contestats</u>

Un cop hem fusionat els dos data sets, veiem que existeixen valors NA o en blanc per a les variables Survived, Fare, Cabin, Embarked i Age.

La variable Age conté 263 valors no contestats, els quals representen el 20.1% del total; la variable Embarked conté 2 valors buits, que representen el 0.15% del total; la variable Fare té un valor en blanc; la variable Survived en conté 468 i, finalment, la variable Cabin conté 1.014 valors buits, que representen el 77.46% dels casos.

Respecte la variable Survived, no és sorprenent que tinguem valors en blanc, ja que per a totes les observacions del data set "test.csv", aquesta no existeix. Ja que és la nostra variable de predicció, no tractarem aquests valors NA.

Pel que fa a la variable Cabin, ja que la majoria de valors es troben en blanc, prescindirem d'ella per a la anàlisi.

Respecte la variable Embarked, substituïm els valors buits pel valor més comú. En canvi, per a la variable Fare, agafarem la mitjana per l'únic valor que falta.

Finalment, 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 dels valors es troben no contestats. En aquest cas, substituirem els valors NA en funció dels passatgers semblants en funció de les altres variables.

Creació de noves variables

Primerament, hem creat una anomenada “FamilySize”, la qual indica la grandària de la família, on el valor mínim és 1, en el cas de que el passatger viatgés sol. Per tal realitzar-la, hem utilitzat 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, hem creat una altra variable, anomenada “Title”, que indica el títol del passatger o tripulant. Aquesta variable la crearem a partir de l’extracció de la primera part del nom (“Name”) del passatger. Aquesta variable l’hem agrupat de 18 possibilitats a 5, per tal de reduir-ne la complexitat.

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 67.

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

Discretització de variables

Hem optat per discretitzar les variables Age i FamilySize per tal de reduir-ne la complexitat. Per tant, hem agrupat els grups d’edat en Children, Young, Adult, Advanced Adult i Elderly. Per altre costat, hem agrupat la mida de la família en Single, Small i Large.

Anàlisi de les dades

Objectius

En aquest apartat, caldrà analitzar la distribució de les variables, la seva normalitat, 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

Distribució Survived

En primer lloc, hem analitzat la distribució de la variable de predicció. És a dir, del total d'observacions, hem comprovat que la majoria no van sobreviure al naufragi. De fet, només un 38.38% dels passatgers i tripulació ho van fer.

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.

Homogeneïtat

En el següent pas, hem comprovat si la variància de l'atribut Age és significativament diferent que la de la variable Survived, a través d'un test de variància (F-Test).

Amb un nivell de significació del 5%, acceptem la hipòtesis nul·la, pel que podem concloure que les dues variables tenen variàncies significativament semblants. A més a més, l'Interval de Confiança es troba per sobre de 1, que reafirma que les variables són semblants.

Matriu de correlacions

A partir d'una matriu de correlacions hem volgut comprovar quines variables tenen una correlació significativa entre elles.

Com a resultat, podem veure que, de forma evident, la correlació més forta es troba entre les variables FamilySize amb Parch i SibSp. No és sorprenent ja que la mida de la família s'ha derivat a partir d'aquestes dues variables.

A més a més, també existeix una correlació inversa entre Age i Survived. Això vol dir que a més jove el passatger, més probabilitats hauria de tenir de supervivència.

Alhora, es detecten correlacions entre Sex i Survived, així com Pclass i Survived. I, també, entre Edat i Fare.

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.

Reducció de la dimensionalitat

Després d'haver analitzat totes les variables de forma detallada, hem decidit reduir la dimensionalitat del data set, així reduint el nombre d'atributs. Del total hem eliminat la variable Cabin, tal i com s'ha explicat en l'apartat de tractament de valors NA o nuls. A més a més, eliminem les variables SibSp i Parche, ja que a través de la mida de la família (FamilySize) ja agrupem la mateixa informació. Pel contrari, sinó, estaríem tenint en compte les mateixes variables de forma duplicada. Finalment, també hem eliminat PassengerId i Name, ja que no aporten cap valor afegit.

Per tant, en conclusió, ens hem quedat amb les següents variables: Age, Survived, PClass, Sex, Fare, FamilySize, Embarked i Title.

Divisió del data set

Abans de crear qualsevol model, hem dividit el data set train en dos, un amb 75% dels casos i, l'altre, amb la resta. El primer l'hem utilitzat per a crear diferents models explicats en el següent apartat, mentre que l'altre per testejar la bondat dels models.

Models: resolució i representació del resultat

Objectius

L'objectiu d'aquest apartat és crear diversos models supervisats per tal de veure quin té una precisió més alta. Aquest serà més tard utilitzat per a realitzar prediccions.

Principalment, hem realitzat tres tipus de models: regressions logístiques, Arbres de decisió i Random Forest.

Resolució

Models de regressió univariables

En aquest punt, hem creat dos models de regressió per a testejar les variables Sex i Embarked en funció de Survived. Hem escollit Sex i Embarked de manera aleatòria i, per tant, hem creat un model amb la variable dependent Survived i cada atribut com a variable independent, 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 92.8% 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 46.6% respecte els del port C.

Aquests dos models senzills ens serveixen per veure que seria beneficiós provar de construir un model de regressió amb totes les variables. Per tant, a continuació, hem creat un model de regressió logística multivariable.

Model de regressió multivariable

Per tant, 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 600, 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 empitjorat lleugerament, amb un resultat no satisfactori. A més a més, hem observat que la variable FamilySize només és significativa en alguns nivells. Per tant, hem optat per realitzar un tercer model, excloent també l'atribut FamilySize.

En aquest tercer model, el valor AIC millora per sobre dels 600 on totes les variables són significatives, al menys a un nivell del 1%. Per tant, dels tres models utilitzats, podem concloure que aquest és el més precís de tots.

Tot i així, hem provat de tornar a executar aquest tercer model però amb la variable Age factoritzada. Per a aquest quart model, el AIC és el més alt dels quatre i totes les variables són significatives. Per tant, dins dels models de regressió logística multivariables, ens quedem amb aquest quart i últim model.

Finalment, hem comprovat la bondat d'aquest quart model, així testejant la predicció a partir del 25% 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 76.2% de precisió. És a dir, en el 76.2% dels casos, el model predirà de forma correcta si el passatger, donades les variables, sobreviuria o no al naufragi. En 116 casos, predirà de forma correcte que el passatger no sobreviuria i, en 54 casos, que sí sobreviuria. Per altre costat, en 32 ocasions predirà que sí sobreviurà quan de fet no ho va fer i, en 21 casos, predirà que no sobreviuria, quan sí ho va fer.

Decision Tree Classification

A continuació, hem entrenat un altre tipus de model, un arbre de decisió, per comprovar si aquest tipus de model pot millorar la precisió de l'últim model entrenat.

Aquest model prediu la probabilitat de supervivència de la següent manera:

- Comprova si el títol del passatger era Mr. Officer. En cas de ser-ho, només un 15% sobreviuria al naufragi. Per tant, el 60% de tots els passatgers i tripulació no sobreviuria.
- En cas de no tenir un títol Mr. Officer, comprova en quin tipus de classe viatjava el passatger. Si era ho feia en tercera classe, el 95% no sobreviuria, que representa el 21% del total de passatgers.
- Si el passatger viatjava en primera o segona classe, tenia un 47% de supervivència, que representa un 19% del total de passatgers. A continuació, comprova si el passatger va pagar un bitllet igual o més car de 23GBP. En cas que sí, la probabilitat de supervivència era del 3% dins d'aquest tipus de passatgers, mentre que la probabilitat seria d'un 5% tenint en compte tots els passatgers.
- Si el passatger va pagar menys de 23GBP pel bitllet i l'edat era igual o superior a 19 anys, el 88% dels passatgers van morir, que representa un 5% del total.
- Per altre costat, si el passatger tenia una edat inferior a 19 anys i va pagar igual o més de 15GBP pel bitllet, tenia una probabilitat de supervivència del 79%, que representa un 2% del total.
- En cas que pagués menys de 15GBP però igual o més de 8GBP pel bitllet, la probabilitat de supervivència és del 64%, que representa un 4% del total. Per contra, si va pagar menys de 8GBP, tenia un 15% de probabilitats de no sobreviure.

En comprovar la bondat del model, veiem que té un nivell de precisió del 78.03%, el qual supera el del model multivariant anterior.

Random Forest Classification

Finalment, hem executat un últim model, basat en Random Forest. Aquest model prediu de forma correcta 378 no sobreviurien al naufragi i 186 que sí ho farien. Per altra banda, en 34 ocasions prediu que no sobreviuria quan sí ho faria i, en 70, prediu que sí sobreviuria quan de fet no seria el cas.

Quan hem comprovat al bondat del model, aquest té una precisió del 80.3%. Per tant, podem concloure que aquest és el millor model de tots per a realitzar prediccions.

Finalment, també ens ha interessat veure quines són les variables amb més influència i, per tant, més rellevants dins del model. Aquestes han estat Title, Sex i Fare.

Prediccions

Objectius

En aquest últim punt, l'objectiu és realitzar una predicció a partir del fitxer "test.csv".

Resolució

En primer lloc, hem realitzat la predicció del passatger número 71. Aquest, tenia un número Id de 962, on el model prediu que el passatger sí que sobreviuria, donades les diferents característiques.

I, finalment, hem aplicat la predicció de tots els passatgers del fitxer "test.csv". El resultat es troba guardat en un fitxer anomenat "test_amb_prediccions.csv".

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