Tipología y ciclo de vida de los datos
Aitor Prieto Moreno

Practica 2: Limpieza y validación de datos

Contenido

Tipología y ciclo de vida de los datos
Aitor Prieto Moreno

Descripción del dataset

El conjunto de datos que se ha seleccionado para la elaboración de esta practica se encuentra en el siguiente enlace:

https://www.kaggle.com/c/titanic/data

Este conjunto de datos contiene información sobre cada uno de los pasajeros del Titanic. Cada registro refleja la información de un pasajero y si pudo sobrevivir o no al desastre del Titanic.

Este dataset esta compuesto por 891 registros y 12 columnas. Los campos que representan cada una de las columnas son los siguientes:

- PassengerId: identificador del pasajero en cuestión.
- survival: indica si el pasajero en cuestión sobrevivió o no. Si el valor es 0 significa que no sobrevivió, pero si el valor es 1 significa que si.
- Name: indica el nombre del pasajero en cuestión.
- pclass: clase socioeconómica del pasajero. Si es igual a 1 es clase alta, si es igual a 2 es clase media y si es clase 3 es clase baja.
- sex: indica el genero del pasajero, male(masculino) o female(femenino).
- Age: numérico que indica la edad del pasajero.
- sibsp: numero de conyugues o hermanos a bordo.
- parch: numero de padres o hijos a bordo. Si se es un niño y este registro contiene el valor 0 sígnica que ha ido acompañado por una niñera.
- ticket: número de ticket(billete) adquirido por este pasajero.
- fare: tarifa del viaje a este pasajero.
- cabin: indica la cabina en la que se aloja el pasajero.
- embarked: indica el puerto donde ha embarcado el pasajero.

Objetivos del análisis

El objetivo de esta práctica será determinar que variables determinan si un pasajero sobrevive o no. Una vez se sepan cuales son estas variables se creará un modelo que permita predecir si un pasajero en concreto puede sobrevivir o no.

Este análisis no creo que pueda ayudar actualmente a mejorar la seguridad, pero creo que a nivel de curiosidad o histórico puede ser interesante. Saber si el motivo de que la gente que sobreviviera era por causas socioeconómicas, por el numero de familiares que se tenía, cuestión física(edad y sexo), o simplemente por si estaba alojado en una habitación en concreto.

Limpieza de datos

Antes de iniciar con la limpieza de datos hace falta analizar el contenido del fichero y la estructura de este:

```
#Lectura del fichero
titanic<-read.csv("D:/Aitor/Escritorio/train.csv",header=TRUE,sep=",")
```

```
> #Dimensiones de los datos
> dim(titanic)
[1] 891  12
```

Como se ha indicado anteriormente se pude observar que contiene 891 registros y 12 columnas. De esta cantidad de registros o columnas después se vera cual es la cantidad real que se utilizara para realizar el análisis, ya que puede ser que se tengan que eliminar columnas o modificar y eliminar registros.

```
> #Estructura de los datos
> str(titanic)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 $
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 2$
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1$
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Aquí se puede observar que variables hay y de que tipo son. Como se ha mencionado anteriormente, puede que más adelante se tenga que transformar o eliminar algunas de estas variables

Integración y datos de interés a analizar

De las variables del dataset se ha decidido eliminar las siguientes:

- PassengerId: identificador del pasajero en cuestión.
- Name: indica el nombre del pasajero en cuestión.
- ticket: número de ticket(billete) adquirido por este pasajero.
- cabin: indica la cabina en la que se aloja el pasajero. Este podría ser un dato interesante de analizar, pero hay demasiadas de ellas y la fuente de datos en ningún momento indica si existe algún tipo de relación entre cada una de ellas.
- fare: indica la tarifa del viaje. No creo que sea relevante si hay un campo que indica la clase socioeconómica.
- embarked: indica el puerto donde ha embarcado el pasajero.

Se ha decidido eliminar estas variables ya que no creo que sean relevantes a la hora de decidir si un pasajero sobrevive o no.

Para eliminar estos datos se ha usado la instrucción select(). Para poder usarla hay que incluir la siguiente librería:

```
#Librerías utilizadas
if(!require(dplyr)){
    install.packages('dplyr', repos='http://cran.us.r-project.org')
    library(dplyr)
}
```

La instrucción se ejecutaría de la siguiente manera:

```
#Eliminación de columnas no útiles
titanic <- select(titanic,-PassengerId,-Name,-Ticket,-Embarked,-Cabin,-Fare)
```

También se ha transformado el campo “Pclass” a factorial ya que ese campo actúa como campo cualitativo.

```
> #Tranformar Pclass a factorial
> titanic$Pclass <- factor(titanic$Pclass)
```

Y el resultado sería el siguiente:

```
> #Nueva estructura de datos
> str(titanic)
'data.frame':   891 obs. of  6 variables:
 $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
```

Ceros y elementos vacíos

En este conjunto de datos no se utiliza el valor 0 para indicar la ausencia de valores en registros. Para indicar esta ausencia se utiliza se deja vacío el campo en concreto.

Para contar el numero de campos vacíos se ha utilizado la siguiente instrucción:

```
> sapply(titanic, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch
       0        0        0      177        0        0
```

Como se pude comprobar se puede observar el campo Age esta vació en 177 registros.

Se ha decidido eliminar estos registros. Se podría haber intentado rellenar esos registros comparando la información entre otros y encontrando similitudes entre sí, el problema es que los campos “sibsp” y “parch” no especifican si el número de familiares es de un tipo en concreto o de otro. Por ejemplo, si el campo “parch” indicara tan solo el número de hijos podría ayudar a la hora de decidir si una persona es adulta o no, el problema es que puede indicar el numero de padres o de hijos.

Para la eliminación de estos registros se ha utilizado la siguiente instrucción:

```
#Eliminación de registros con Age vacios
titanic <- titanic[!is.na(titanic$Age),]
```

Se puede comprobar que el numero de campos vacíos actualmente es 0:

```
> sapply(titanic, function(x) sum(is.na(x)))
Survived   Pclass      Sex      Age    SibSp    Parch
       0        0        0        0        0        0
```

Valores extremos

Para encontrar que valores destacan por su falta de congruencia delante del resto de datos se ha usado se ha usado la instrucción boxplot.stats()$out sobre las variables numericas:

```
> #Identificacion de valores extremos
> boxplot.stats(titanic$Survived)$out
integer(0)
> boxplot.stats(titanic$Age)$out
[1] 66.0 71.0 70.5 71.0 80.0 70.0 70.0 74.0
> boxplot.stats(titanic$SibSp)$out
 [1] 3 4 3 3 4 5 3 4 5 3 3 4 4 4 4 4 4 4 4 3 3 5 5 4 4 3 3 5 4 3 4 4 3 4 4
> boxplot.stats(titanic$Parch)$out
 [1] 5 5 3 4 4 3 4 4 5 5 6 3 3 3 5
```

Como se pude comprobar no hace falta preocuparse por esos valores, ya que es posible que en el titanic hubieran pasajeros con esas edades, con esa cantidad de hermanos o esa cantidad de hijos (sobre todo en aquella época).

Exportación de datos

Ahora que ya se ha analizado los datos origen, se han integrado y validado, se generada un fichero que contendrá estos datos resultantes:

```
#Exportar datos
write.csv(titanic,"clean_titanic.csv")
```

Análisis de datos

Selección de los grupos de datos a analizar/comprarar

A partir de los diferentes campos cualitativos con sus respectivos valores se han creado diferentes grupos, los cuales algunos han sido utilizados para realizar diferentes pruebas estadísticas.

```
#Agrupacion por clase
titanic.Grupo1 <-titanic[titanic$Pclass.type == "1"]
titanic.Grupo2 <-titanic[titanic$Pclass.type == "2"]
titanic.Grupo3 <-titanic[titanic$Pclass.type == "3"]

#Agrupacion por Sex
titanic.Male <-titanic[titanic$Sex.type == "male"]
titanic.Female <-titanic[titanic$Sex.type == "female"]
```

Comprobación de la normalidad i homogeneidad de la variancia

Para comprobar si los valores de las variables cuantitativas forman parte de una distribución normal se ha utilizado el test Kolmogorov-Smirnov(concretamente una modificación llamada Lillifors). Para hacer uso de este test se ha tenido que instalar el siguiente paquete:

```
if(!require(nortest)){
    install.packages('nortest', repos='http://cran.us.r-project.org')
    library(nortest)
}
```

Una vez instalado se ha ejecutado de la siguiente manera:

```
> #Revisar si las variables cuantitativas estan normalizadas
> #Age
> lillie.test(titanic$Age)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  titanic$Age
D = 0.064567, p-value = 1.862e-07

> #SibSP
> lillie.test(titanic$SibSp)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  titanic$SibSp
D = 0.36896, p-value < 2.2e-16

> #Parch
> lillie.test(titanic$Parch)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  titanic$Parch
D = 0.4231, p-value < 2.2e-16
```

Como se puede comprobar en los 3 resultados el p-valor es inferior a al coeficiente 0.05. Lo que indica que ninguna de las variables sigue una distribución normal.

Una vez comprobada la normalidad de las variables, se comprobará la homogeneidad de varianza. Para ello se hará uso del test Fligner-Killeen utilizando los grupos conformados a partir de la variable “Sex” y la variable “Survived”.

```
> #Test Fligner-Killeen
> fligner.test(Survived ~ Sex, data=titanic )

        Fligner-Killeen test of homogeneity of variances

data:  Survived by Sex
Fligner-Killeen:med chi-squared = 1.5357, df = 1, p-value = 0.2153
```

Como se puede observar el valor de p-valor es superior a 0.05 eso nos permite validar la homogeneidad de las hipótesis de las dos muestras.

Representación de resultados a partir de tablas y graficas

Variables cuantitativas que influyen a lo hora de determinar la supervivencia

En este apartado se determinará que variables cuantitativas influyen mas a la hora de determinar la supervivencia. Para ello se hará un análisis de correlación sobre esas variables i la variable de supervivencia. Ya que estos datos cuantitativos no siguen una distribución normal se hará uso del test de correlación de Spearman.

```
> #Calculo del coeficiente de correlacion de Kendall para cada variable cuantitativa respecto al campo Survived
> #Age
> spearman_test1 = cor.test(titanic$Age, titanic$Survived, method = "kendall")
> #SibSP
> spearman_test2 = cor.test(titanic$SibSp, titanic$Survived, method = "kendall")
> #Parch
> spearman_test3 = cor.test(titanic$Parch, titanic$Survived, method = "kendall")
>
> #Tabla con las correlaciones de spearman
> tabla.corr <- matrix(c("Age",spearman_test1$estimate,spearman_test1$p.value,
+                        "SibSP",spearman_test2$estimate,spearman_test2$p.value,
+                        "Parch",spearman_test3$estimate,spearman_test3$p.value),ncol=3,byrow = TRUE)
> colnames(tabla.corr) <- c("campo","estimate","p-value")
> print(tabla.corr)
     campo    estimate              p-value
[1,] "Age"    "-0.0433850545172538" "0.160437462522075"
[2,] "SibSP"  "0.0707754372192448"  "0.0504933648259012"
[3,] "Parch"  "0.150952419261913"   "2.94886469958888e-05"
```

Una vez se obtiene los resultados, para determinar si una de estas variables esta correlacionadas con el campo supervivencia se debe comprobar la distancia la cual se encuentra el valor estimado respecto a los valores -1 y 1. Esta cercanía refleja la correlación que hay entre una de esas variables y el campo supervivencia. En este caso se puede comprobar que el campo mas cercano es el Parch (número de padre o hijos a bordo).

Modelo de regresión lineal realizado

Una vez encontrado el nivel de correlación de cada una de las variables cuantitativas se procederán a crear diversos modelos de regresión lineal que harán uso de estas variables junto a las variables cualitativas. Para generar cada uno de los modelos se hará uso de la instrucción lm().

```
> #Modelos de regresión lineal
> #Regresores cuantitativos
> Age=titanic$Age
> SibSP=titanic$SibSp
> Parch=titanic$Parch
>
> #Regresores cualitativos
> Pclass = titanic$Pclass
> Sex = titanic$Sex
>
> #Variable a predecir
> Survived = titanic$Survived
>
> #Definicion de modelos
> modelo1 <- lm (Survived ~ Pclass + Sex + Age + SibSP + Parch, data = titanic )
> modelo2 <- lm (Survived ~ Pclass + Sex + Age + SibSP , data = titanic )
> modelo3 <- lm (Survived ~ Pclass + Sex + Age + Parch, data = titanic )
> modelo4 <- lm (Survived ~ Pclass + Sex , data = titanic )
> modelo5 <- lm (Survived ~ Sex + Age , data = titanic )
```

Una vez generado cada uno de los modelos se analizarán cuál de estos es más eficiente. Para ello se analizará su coeficiente de determinación.

```
> #Tabla con los coeficientes de determinación de cada modelo
> tabla.coe <- matrix(c("Modelo1",summary(modelo1)$r.squared,
+                       "Modelo2",summary(modelo2)$r.squared,
+                       "Modelo3",summary(modelo3)$r.squared,
+                       "Modelo4",summary(modelo4)$r.squared,
+                       "Modelo5",summary(modelo5)$r.squared),ncol=2,byrow = TRUE)
> colnames(tabla.coe) <- c("modelo","coeficiente determinacion")
> print(tabla.coe)
     modelo    coeficiente determinacion
[1,] "Modelo1" "0.40002263289444"
[2,] "Modelo2" "0.399822242850968"
[3,] "Modelo3" "0.39235395804175"
[4,] "Modelo4" "0.368343680369509"
[5,] "Modelo5" "0.291066981842692"
```

Como se puede comprobar el coeficiente mas alto es del modelo 1. Si observamos detalladamente podemos observar que entre los 4 primeros modelos no hay mucha diferencia entre los coeficientes de determinación. Pero en el momento que no se utiliza una variable cualitativa este coeficiente baja considerablemente. En concreto como se pude observar la variable “Sex” tiene un gran peso a la hora de determinar la supervivencia.

Se ha utilizado el dato para realizar una predicción sobre el modelo 1:

```
> #Datos para probar prediccion
> Superviviente1 <- data.frame(Pclass = "1", Sex = "male", Age = 14, SibSP = 3, Parch = 2)
>
> #Predicion a partir del modelo1
> predict(modelo1, Superviviente1)
        1
0.4450757
```

Como se puede observar, una persona con estas características tiene mas posibilidades de sobrevivir que de no sobrevivir.

Ficheros utilizados

Los ficheros que se han utilizado para la elaboración de esta práctica son los siguientes:

- titanic.R : script realizado para la elaboración de la práctica.
- Practica2.pdf : informe de la practica 2.
- train.csv: datos originales descargados en https://www.kaggle.com/c/titanic/data
- clean_titanic.csv : datos resultantes después de realizar los procesos de limpieza e integración.

Conclusiones

Después de realizar todos los pasos descritos en la práctica he llegado a la conclusión de que la variable que influye mas sobre el campo supervivencia es la que describe el genero de una persona. Aun así, hace falta mencionar que aunque genere una diferencia considerable sobre el índice de determinación respecto las otras este aun sigue siendo bajo. Eso me lleva a pensar a que tal vez uno de los campos eliminados inicialmente en el dataset (cabina o puerto de embarque) si tenga mas repercusión de la que creía inicialmente. Aun así, considero que todavía falta algún tipo de información en la pagina de la fuente de datos donde se describa más detalladamente esos campos.