A1 - Preproceso de datos

Enunciado

Semestre 2021.1

Índice

Introducción

En esta actividad realizaremos el preprocesado de un fichero de datos que contiene información de una muestra de indemnizaciones otorgadas por una compañía de seguros, en función del tiempo de baja laboral del trabajador. El conjunto de datos tiene 54000 registros y 15 variables. Los datos se han extraído y modificado parcialmente (por motivos académicos) de la base de datos disponible en la plataforma Kaggle: https://www.kaggle.com/c/actuarial-loss-estimation.

Las variables del fichero de datos (**train3.csv**) son:

- ClaimNumber: Identificador de la póliza.
- DateTimeOfAccident: Fecha del accidente.
- DateReported: Fecha que se comunica a la compañía y se abre el expediente.
- Age: Edad del trabajador.
- Gender: Sexo.
- MaritalStatus: Estado civil, (M)arried, (S)ingle, (U)nknown, (W)idowed, (D)ivorced.
- DependentChildren: Número de hijos dependientes.
- DependentsOther: Número de dependientes excluyendo hijos.
- WeeklyWages: Salario semanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWorkedPerWeek: Número horas por semana.
- DaysWorkedPerWeek: Número de días por semana.
- ClaimDescription: Descripción siniestros.
- InitialIncurredClaimCost: Estimación inicial del coste realizado por la compañía.
- UltimateIncurredClaimCost: Coste total pagado por siniestro.

El objetivo de esta actividad es preparar el fichero para su posterior análisis. Para ello, se examinará el fichero para detectar y corregir posibles errores, inconsistencias y valores perdidos. Además, se presentará una breve estadística descriptiva.

Criterios de verificación y de normalización de las variables:

A continuación se muestran los criterios con los que deben limpiarse los datos del conjunto:

1. Verificar si hay registros duplicados con el valor ClaimNumber. En caso de duplicación, se deben asignar nuevos códigos no repetidos.
2. Los valores posibles en la variable MaritalStatus son: M (married), S (single), U (unknown), D (divorced), W (widowed).
3. Los valores posibles en la variable Género son: F (femenino), M (masculino), U (unknown).
4. En los datos numéricos, el símbolo de separador decimal es el punto y no la coma.
5. Las variables InititalIncurredClaimsCost y UltimateIncurredClaimsCost se debe expresar como variables numéricas (de tipo entero), en unidades (no en miles) y sin decimales.
6. Las variables Age y DaysWorkedPerWeek deben ser de tipo entero, sin decimales.
7. Las variables WeeklyWages y HoursWorkedPerWeek deben expresarse como variables numéricas y si es necesario, con decimales.

Aspectos importantes a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado. Si un apartado no se responde, se puede dejar en blanco para respetar el orden y numeración del documento.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos. Es decir, se debe usar la terminología propia de la estadística.
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. El documento debe estar bien estructurado y ser conciso.

Para realizar el preproceso del fichero, seguir los pasos que se indican a continuación.

---

1. Carga del archivo

Cargar el archivo de datos y examinar el tipo de datos con los que R ha interpretado cada variable. Examinar también los valores resumen de cada tipo de variable.

2. Duplicación de códigos

Verificad la consistencia en la variable ClaimNumber. Si existen registros duplicados, asignad un nuevo código para evitar códigos duplicados. El nuevo código debe ser un valor no usado (valores superiores al máximo valor numérico contenido en ClaimNumber). Conservad el mismo formato que el resto de códigos, con "WC" delante de la secuencia numérica. Podéis usar la función **duplicated** de R para detectar los duplicados.

3. Nombres de las variables

Simplificad el nombre de algunas variables para hacer más fácil su manejo. En concreto, cambiad el nombre de las variables:

- InitialIncurredClaimCost por IniCost
- UltimateIncurredClaimCost por UltCost
- HoursWorkedPerWeek por HoursWeek
- DaysWorkedPerWeek por DaysWeek

En el resto del enunciado, usaremos los nuevos nombres de estas variables.

4. Normalización de los datos cualitativos

4.1. Marital Status

Los valores posibles en marital status son: M (married), S (single), U (unknown), D (divorced), W (widowed). Los valores nulos corresponden a unknown. Si hay valores perdidos (vacíos), modificad estos registros vacíos para que en su lugar aparezca U (unknown). Revisad también la consistencia del resto de valores de esta variable.

4.2. Género

Revisad la consistencia de los valores de la variable género y realizad las modificaciones oportunas.

5. Normalización de los datos cuantitativos

Inspeccionar los valores de los datos cuantitativos y realizar las normalizaciones oportunas siguiendo los criterios especificados anteriormente. Estas normalizaciones tienen como objetivo uniformizar los formatos. Si hay valores perdidos o valores extremos, se tratarán más adelante.

Al realizar estas normalizaciones, se debe demostrar que la normalización sobre cada variable ha dado el resultado esperado. Por lo tanto, se recomienda mostrar un fragmento del archivo de datos resultante. Para evitar mostrar todo el conjunto de datos, se puede mostrar una parte del mismo, con las funciones **head** y/o **tail**.

Seguid el orden de los apartados.

5.1. IniCost y UltCost

Revisad el formato de estas variables según el criterio indicado más arriba. Si existen valores extremos, se tratarán más adelante.

5.2. Edad

Revisad el formato de la variable Age y realizad las transformaciones oportunas según los criterios especificados anteriormente. Si existen valores extremos, se tratarán más adelante.

5.3. WeeklyWages, HoursWeek, DaysWeek

Revisad el formato de estas variables y realizad las transformaciones oportunas según los criterios especificados anteriormente. Visualizad la distribución de los valores en estas tres variables, usando el tipo de gráfico adecuado. Si existen valores extremos, se tratarán más adelante.

6. Valores atípicos

Revisad si hay valores atípicos en las variables Age, WeeklyWages, HoursWeek y DaysWeek. Si se trata de un valor anómalo, es decir anormalmente alto o bajo, substituir su valor por NA y posteriormente se imputará.

7. Imputación de valores

Buscad si existen valores perdidos en las variables cuantitativas Age, WeeklyWages, HoursWeek, IniCost y UltCost

En caso de valores perdidos, aplicad el proceso siguiente:

- Para Age, aplicad imputación por la media aritmética.
- En el resto de variables, aplicad imputación por vecinos más cercanos, usando la distancia de Gower, considerando en el cómputo de los vecinos más cercanos el resto de variables cuantitativas mencionadas en este apartado. Además, considerad que la imputación debe hacerse con registros del mismo género. Per exemple, si un registro a imputar es de género "M", se debe realizar la imputación usando las variables cuantitativas de los registros de género "M". Para realizar esta imputación, podéis usar la función "kNN" de la librería VIM.

Mostrad que la imputación se ha realizado correctamente, mostrando el resultado de los datos afectados por la imputación.

8. Preparación de los datos

8.1. Tiempo de abertura del expediente

Calculad el tiempo que se tarda en abrir el expediente desde el suceso del accidente, a partir de las variables DateOfTimeAccident y DateReported. Para ello, debéis convertir las variables en formato fecha (Date) y realizar posteriormente el cálculo. Guardad la información en una variable Time del conjunto de datos.

8.2. Diferencia entre IniCost y UltCost

Calculad la diferencia entre el coste final y el coste inicial estimado. Añadid una variable "DifCost" al conjunto de datos y visualizad su distribución con un gráfico adecuado.

9. Estudio descriptivo

9.1. Funciones de media robustas

Implementad una función en R que, dado un vector con datos numéricos, calcule la media recortada y la media Winsor. Estas funciones se deben definir como sigue:

```
media.recortada <- function( x, perc=0.05){}

media.winsor( x, perc=0.05){}
```

donde x es el vector de datos y perc la fracción de los datos a recortar (por defecto, 0.05). Implementad estas funciones en R y comprobad que funcionan correctamente.

9.2. Estudio descriptivo de las variables cuantitativas

Realizad un estudio descriptivo de las variables cuantitativas Age, WeeklyWages, DaysWeek, HoursWeek, IniCost, UltCost.

Para ello, preparad una tabla con varias medidas de tedencia central y dispersión, robustas y no robustas. Usad, entre otras, las funciones del apartado anterior. Presentad, asimismo gráficos donde se visualice la distribución de los valores de estas variables cuantitativas.

10. Archivo final

Una vez realizado el preprocesamiento sobre el archivo, copiad el resultado de los datos en un archivo llamado "train_clean.csv".

11. Evaluación de la actividad

- Apartados 1,2 (20 %)
- Apartados 3,4 (10 %)
- Apartado 5 (10 %)
- Apartado 6 (10 %)
- Apartado 7 (10 %)
- Apartado 8 (10 %)
- Apartados 9,10 (20 %)
- Calidad del informe dinámico (calidad del código, formato y estructura del documento, concisión y precisión en las respuestas) (10 %)