Descripción del proceso del tratamiento de datos

Los datos se pueden descargar de:

https://www.kaggle.com/c/liberty-mutual-fire-peril/download/train.csv.zip

https://www.kaggle.com/c/liberty-mutual-fire-peril/download/test.csv.zip

Tratamiento del fichero train

El tratamiento del fichero de datos se va a realizar en Python.

El fichero se carga como un dataframe al que llamaremos df:

```
df = pd.read_csv("train.csv",sep=",")
```

Se comprueba que el fichero no tenga registros duplicados:

```
df = df.drop_duplicates()
```

Se comprueban los tipos de las variables del dataframe:

```
df.dtypes
id              int64
target          float64
var1            object
var2            object
var3            object
.....
weatherVar234   float64
weatherVar235   float64
weatherVar236   float64
```

Comprobamos que la variable identificador no contiene nulos y se transforma a string.

df.id.isnull().value_counts()

False 452061

df.id = df.id.astype(str)

Comprobamos que la variable target no contiene nulos y analizamos sus valores

df.target.isnull().value_counts()

False 452061

df.target.value_counts()

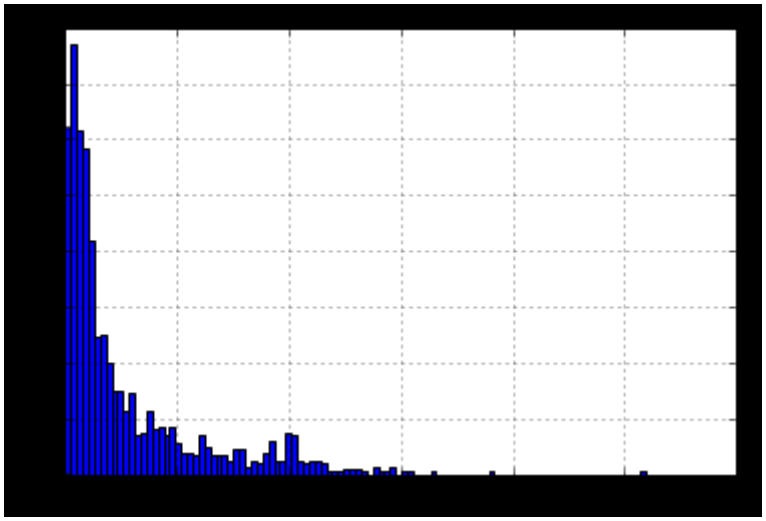
0.0 450873

……..

25.920137918000002 1

El 99% de los valores de target son ceros el resto varía entre >0 y 25.92

Histograma de los valores positivos de target



Analizamos las restantes variables

Obtenemos la lista de campos del dataframe df y en particular las categóricas:

fieldslist = df.columns.tolist()

Lista de variables categóricas:

catfieldslist = fieldslist[2:11]

Analizamos los distintos valores de éstas variables teniendo en cuenta que, según la documentación, los valores 'Z' representan valores nulos

var1

['Z' '3' '2' '4' '1' '5']

var2

['Z' 'A' 'B' 'C']

var3

['Z' '4' '3' '5' '2' '6' '1']

var4

['N1' 'C1' 'J3' 'H1' 'A1' 'H2' 'E3' 'M1' 'R3' 'I1' 'J4' 'D3' 'R6' 'O2' 'F1'

'R5' 'Z' 'H3' 'D1' 'P1' 'J5' 'J2' 'E2' 'S1' 'E1' 'L1' 'G1' 'D2' 'O1' 'B1'

'G2' 'E6' 'K1' 'D4' 'E5' 'J1' 'E4' 'R7' 'R2' 'R8' 'J6' 'R1' 'R4']

var5

['Z' 'B' 'D' 'E' 'C' 'A' 'F']

var6

['Z' 'B' 'A' 'C']

var7

['3' '2' '4' '7' '5' '1' '6' '8' 'Z' 2 6 5 4 3 7 1 8]

var8

['1' '2' '4' '3' '6' 'Z' '5']

var9

['A' 'Z' 'B']

Los valores Z representan los nulos y se tratará como una categoría mas.

La variable var7 presenta valores númericos y de texto, asumiendo que son equivalentes, pondremos todos los valores como texto

df.var7 = df.var7.astype(str)

Variables Numéricas relativas a la poliza

Comprobación si existen valores nulos en las variables en las variables: var10 a var17.

```
fieldslist = df.columns.tolist()
numericlist = fieldslist[11:19]
for i in numericlist:
    print(i)
    namefield = "df."+i
    print(df[i].isnull().unique())
    print("")
    df[i]
```

var10

[False]

var11

[False]

var12

[False][True]

var13

[False]

var14

[False][True]

var15

[False][True]

var16

[False][True]

var17

[False]

Las variables var12, var14, var15, var16 presentan valores nulos que se sustituyen por la moda de cada variable

```
vr = ['var12','var14','var15','var16' ]

for i in vr:
    mode = df[i].mode()[[0][0]]
    df[i] = df[i].fillna(mode)
```

Variables Numéricas relativas a criminalidad

Comprobación de valores nulos en éstas variables

Análogamente se comprueba que todas las variables “crime” presentan valores nulos que se sustituirán por la moda en cada caso.

Variables Numéricas geodemográficas y meteorológicas

Comprobación de valores nulos en éstas variables

Análogamente se comprueba que todas las variables “weather” presentan valores nulos que se sustituirán por la moda en cada caso.

Conversión de variables categóricas en dummies

Convertimos las variables categóricas en dummies

```
dfcat = pd.DataFrame(df,columns = catfieldslist)
```

```
dfcatdummies = pd.get_dummies(dfcat)
```

Eliminamos del dataset df las variables categóricas

```
for i in catfieldslist:
    df = df.drop(i,axis=1)
```

Unión del dataframe original con el de las variables dummies

```
df = df.join(dfcatdummies)
```

Escritura en disco del dataframe definitivo:

```
df.to_csv("train_clean.csv",sep=";")
```

Tratamiento del fichero test

Será totalmente análogo, solo que los valores omitidos se sustituirán por la moda de la variable del conjunto de train.