Reclamaciones de seguros: detección de fraudes

sonia rodriguez del cerro garcia

diciembre de 2021

Seleccion del conjunto de datos

Utilizando un conjunto de datos proporcionado por Derrick Mwiti publicados en 31 Oct 2018 https://github.com/mwitiderrick aplicaremos visualizaciones para comprender mejor y analizar estadísticamente los datos.

Para la elaboración de la base de datos se ha decidido extraer una muestra del total de siniestros ocurridos en un período concreto de tiempo de 1.000 expedientes.

Las variables que contiene el conjunto de datos son las siguientes:

- months_as_customer :meses como cliente
- age :edad
- policy_number :número de póliza
- policy_bind_date :fecha de alta de la póliza
- policy_state :estado de la póliza
- policy_csl :cobertura de responsabilidad y por daños a la propiedad
- policy_deductable :importe deducible
- policy_annual_premium :prima anual
- umbrella_limit :límite general
- insured_zip :codigo postal del asegurado
- insured_sex :sexo del asegurado
- insured_education_level :nivel educativo del asegurado
- insured_occupation :ocupacion del asegurado
- insured_hobbies :hobbies del asegurado
- insured_relationship :relación con el asegurado
- capital-gains :ganacias del capital
- capital-loss :perdidas de capital
- incident_date :fecha del incidente
- incident_type :tipo de incidente
- collision_type :tipo de colisión

- incident_severity :gravedad del incidente
- authorities_contacted :autoridades contactadas
- incident_state :estado donde ocurre el incidente
- incident_city :ciudad donde ocurre el incidente
- incident_location :dirección donde ocurre el incidente
- incident_hour_of_the_day :hora del día donde ocurre el incidente
- number_of_vehicles_involved:número de vehículos implicados en el incidente
- property_damage :daño a la propiedad
- bodily_injuries :lesiones corporales
- witnesses :testigos
- police_report_available :informe policial disponible
- total_claim_amount cantidad:total de la reclamación
- injury_claim :reclamo por lesiones
- property_claim :reclamo de propiedad
- vehicle_claim :reclamo de vehículo
- auto_make :marca del vehículo
- auto_model :modelo del vehículo
- auto_year :año del vehículo
- fraud_reported :se trata de un fraude (yes/no)

```
# Cargamos los paquetes R que vamos a usar
library(dplyr)
library(tidyr)
library(gmodels)
library (Hmisc)
# Cargamos los datos
datos <- read.csv('insurance_claims.csv')
```

Dimensión de la base de datos.

En este caso tenemos 39 variables y 1.000 observaciones

```
dim(datos)

## [1] 1000   39
```

Descripcion de las variables

Tipo de cada variable

La base de datos recopilada comprende información relativa a características propias de los asegurados, variables sociodemográficas e información referente a siniestros acaecidos en el pasado.

Los atributos segun su tipo de datos se dividen en dos tipos 18 variables numericas (dbl, int) y el resto cadenas. Podemos convertir las variables de texto a factores (variables con categorías). A continuación, podemos ver la distribución.

```
glimpse(datos)
```

```
## Observations: 1,000
## Variables: 39
## $ months_as_customer          <int> 328, 228, 134, 256, 228, 256, 137,
165,...
## $ age                         <int> 48, 42, 29, 41, 44, 39, 34, 37,
33, 42,...
## $ policy_number               <int> 521585, 342868, 687698, 227811,
367455,...
## $ policy_bind_date            <fct> 2014-10-17, 2006-06-27, 2000-09-
06, 199...
## $ policy_state                <fct> OH, IN, OH, IL, IL, OH, IN, IL,
IL, IL,...
## $ policy_csl                  <fct> 250/500, 250/500, 100/300,
250/500, 500...
## $ policy_deductable           <int> 1000, 2000, 2000, 2000, 1000,
1000, 100...
## $ policy_annual_premium       <dbl> 1406.91, 1197.22, 1413.14,
1415.74, 158...
## $ umbrella_limit              <int> 0, 5000000, 5000000, 6000000,
6000000, ...
## $ insured_zip                 <int> 466132, 468176, 430632, 608117,
610706,...
## $ insured_sex                 <fct> MALE, MALE, FEMALE, FEMALE, MALE,
FEMAL...
## $ insured_education_level     <fct> MD, MD, PhD, PhD, Associate, PhD,
PhD, ...
## $ insured_occupation          <fct> craft-repair, machine-op-inspct,
sales,...
## $ insured_hobbies             <fct> sleeping, reading, board-games,
board-g...
## $ insured_relationship        <fct> husband, other-relative, own-
child, unm...
## $ capital.gains               <int> 53300, 0, 35100, 48900, 66000, 0,
0, 0,...
## $ capital.loss                <int> 0, 0, 0, -62400, -46000, 0, -
77000, 0, ...
## $ incident_date               <fct> 2015-01-25, 2015-01-21, 2015-02-
22, 201...
## $ incident_type               <fct> Single Vehicle Collision, Vehicle
Theft...
## $ collision_type              <fct> Side Collision, ?, Rear Collision,
Fron...
## $ incident_severity           <fct> Major Damage, Minor Damage, Minor
Damag...
```

```
## $ authorities_contacted       <fct> Police, Police, Police, Police,
None, F...
## $ incident_state              <fct> SC, VA, NY, OH, NY, SC, NY, VA,
WV, NC,...
## $ incident_city               <fct> Columbus, Riverwood, Columbus,
Arlingto...
## $ incident_location           <fct> 9935 4th Drive, 6608 MLK Hwy, 7121
Fran...
## $ incident_hour_of_the_day    <int> 5, 8, 7, 5, 20, 19, 0, 23, 21, 14,
22, ...
## $ number_of_vehicles_involved <int> 1, 1, 3, 1, 1, 3, 3, 3, 1, 1, 1,
3, 1, ...
## $ property_damage             <fct> YES, ?, NO, ?, NO, NO, ?, ?, NO,
NO, YE...
## $ bodily_injuries             <int> 1, 0, 2, 1, 0, 0, 0, 2, 1, 2, 2,
1, 1, ...
## $ witnesses                   <int> 2, 0, 3, 2, 1, 2, 0, 2, 1, 1, 2,
2, 0, ...
## $ police_report_available     <fct> YES, ?, NO, NO, NO, NO, ?, YES,
YES, ?,...
## $ total_claim_amount          <int> 71610, 5070, 34650, 63400, 6500,
64100,...
## $ injury_claim                <int> 6510, 780, 7700, 6340, 1300, 6410,
2145...
## $ property_claim              <int> 13020, 780, 3850, 6340, 650, 6410,
7150...
## $ vehicle_claim               <int> 52080, 3510, 23100, 50720, 4550,
51280,...
## $ auto_make                   <fct> Saab, Mercedes, Dodge, Chevrolet,
Accur...
## $ auto_model                  <fct> 92x, E400, RAM, Tahoe, RSX, 95,
Pathfin...
## $ auto_year                   <int> 2004, 2007, 2007, 2014, 2009,
2003, 201...
## $ fraud_reported              <fct> Y, Y, N, Y, N, Y, N, N, N, N, N,
N, N, ...
```

Valores vacios:

```
colSums(is.na(datos))
```

```
##          months_as_customer                         age
##                           0                           0
##               policy_number             policy_bind_date
##                           0                           0
##                policy_state                  policy_csl
##                           0                           0
##           policy_deductable       policy_annual_premium
##                           0                           0
##              umbrella_limit                  insured_zip
##                           0                           0
```

```
##                 insured_sex     insured_education_level
##                           0                           0
##          insured_occupation             insured_hobbies
##                           0                           0
##        insured_relationship               capital.gains
##                           0                           0
##                capital.loss               incident_date
##                           0                           0
##               incident_type              collision_type
##                           0                           0
##           incident_severity       authorities_contacted
##                           0                           0
##              incident_state               incident_city
##                           0                           0
##           incident_location    incident_hour_of_the_day
##                           0                           0
## number_of_vehicles_involved             property_damage
##                           0                           0
##             bodily_injuries                   witnesses
##                           0                           0
##     police_report_available          total_claim_amount
##                           0                           0
##                injury_claim              property_claim
##                           0                           0
##               vehicle_claim                  auto_make
##                           0                           0
##                  auto_model                  auto_year
##                           0                           0
##              fraud_reported
##                           0
```

Variables numericas:

Observamos que para las variables cuantitativas la función summary() proporciona una serie de estadísticos descriptivos relacionados con la posición de la variable (media, mediana, máximo, mínimo,.).

```
estadisticas <- datos %>% select(-policy_bind_date, -policy_state, -
policy_csl, -insured_sex, -insured_education_level, -insured_occupation,
-insured_hobbies, -insured_relationship, -incident_date, -incident_type,
-collision_type, -incident_severity, -authorities_contacted, -
incident_state, -incident_city, -incident_location, -property_damage, -
police_report_available, -auto_make, -auto_model, -fraud_reported) %>%
describe()
estadisticas
```

```
## .
##
##  18  Variables      1000  Observations
## ---------------------------------------------------------------------------
```

```
----------
## months_as_customer
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      391        1      204    130.8     28.9
58.9
##      .25      .50      .75      .90      .95
##    115.8    199.5    276.2    371.0    429.0
##
## lowest :   0   1   2   3   4, highest: 473 475 476 478 479
## ------------------------------------------------------------------------
----------
## age
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0       46    0.999    38.95    10.32       26
28
##      .25      .50      .75      .90      .95
##       32       38       44       53       57
##
## lowest : 19 20 21 22 23, highest: 60 61 62 63 64
## ------------------------------------------------------------------------
----------
## policy_number
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0     1000        1   546239   296820   143970
185105
##      .25      .50      .75      .90      .95
##   335980   533135   759100   914161   954279
##
## lowest : 100804 101421 104594 106186 106873, highest: 996253 996850
998192 998865 999435
## ------------------------------------------------------------------------
----------
## policy_deductable
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        3    0.888     1136    651.2
##
## Value        500  1000  2000
## Frequency    342   351   307
## Proportion 0.342 0.351 0.307
## ------------------------------------------------------------------------
----------
## policy_annual_premium
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      991        1     1256    275.2    855.1
953.9
##      .25      .50      .75      .90      .95
```

```
##   1089.6   1257.2   1415.7   1564.7   1653.4
##
## lowest :  433.33  484.67  538.17  566.11  617.11
## highest: 1922.84 1927.87 1935.85 1969.63 2047.59
## ------------------------------------------------------------------------
----------
## umbrella_limit
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0       11    0.491  1101000  1830577    0e+00
0e+00
##      .25      .50      .75      .90      .95
##    0e+00    0e+00    0e+00    6e+06    6e+06
##
## Value      -1e+06  0e+00  2e+06  3e+06  4e+06  5e+06  6e+06  7e+06
8e+06
## Frequency       1    798      3     12     39     46     57     29
8
## Proportion  0.001  0.798  0.003  0.012  0.039  0.046  0.057  0.029
0.008
##
## Value       9e+06  1e+07
## Frequency       5      2
## Proportion  0.005  0.002
## ------------------------------------------------------------------------
----------
## insured_zip
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      995        1   501214    73341   433274
437419
##      .25      .50      .75      .90      .95
##   448405   466446   603251   614000   617463
##
## lowest : 430104 430141 430232 430380 430567, highest: 620737 620757
620819 620869 620962
## ------------------------------------------------------------------------
----------
## capital.gains
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      338    0.869    25126    29885        0
0
##      .25      .50      .75      .90      .95
##        0        0    51025    64420    70615
##
## lowest :      0    800  10000  11000  12100, highest:  90700  91900
94800  98800 100500
## ------------------------------------------------------------------------
----------
```

```
## capital.loss
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      354    0.893   -26794    30540   -72305   -
65510
##      .25      .50      .75      .90      .95
##   -51500   -23250        0        0        0
##
## lowest : -111100  -93600  -91400  -91200  -90600
## highest:  -10600   -8500   -6300   -5700       0
## ------------------------------------------------------------------------
----------
## incident_hour_of_the_day
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0       24    0.998    11.64    8.021        0
2
##      .25      .50      .75      .90      .95
##        6       12       17       21       23
##
## lowest :  0  1  2  3  4, highest: 19 20 21 22 23
## ------------------------------------------------------------------------
----------
## number_of_vehicles_involved
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        4    0.758    1.839    1.023
##
## Value          1     2     3     4
## Frequency    581    30   358    31
## Proportion 0.581 0.030 0.358 0.031
## ------------------------------------------------------------------------
----------
## bodily_injuries
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        3    0.889    0.992   0.8932
##
## Value          0     1     2
## Frequency    340   328   332
## Proportion 0.340 0.328 0.332
## ------------------------------------------------------------------------
----------
## witnesses
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        4    0.937    1.487    1.243
##
## Value          0     1     2     3
## Frequency    249   258   250   243
## Proportion 0.249 0.258 0.250 0.243
## ------------------------------------------------------------------------
----------
```

```
## total_claim_amount
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      763        1    52762    29128     4320
5756
##      .25      .50      .75      .90      .95
##    41813    58055    70593    81364    88413
##
## lowest :    100   1920   2160   2250   2400, highest: 107900 108030
108480 112320 114920
## ------------------------------------------------------------------------
----------
## injury_claim
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      638        1     7433     5563      450
639
##      .25      .50      .75      .90      .95
##     4295     6775    11305    14380    15662
##
## lowest :     0    10   220   250   280, highest: 18520 19020 20700
21330 21450
## ------------------------------------------------------------------------
----------
## property_claim
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      626        1     7400     5452      450
650
##      .25      .50      .75      .90      .95
##     4445     6750    10885    14142    15540
##
## lowest :     0    20   240   250   260, highest: 21240 21580 21630
21810 23670
## ------------------------------------------------------------------------
----------
## vehicle_claim
##        n  missing distinct     Info     Mean      Gmd      .05
.10
##     1000        0      726        1    37929    20802     3274
4157
##      .25      .50      .75      .90      .95
##    30293    42100    50823    58728    63094
##
## lowest :    70  1440  1680  1750  1760, highest: 76000 76400 77670
77760 79560
## ------------------------------------------------------------------------
----------
## auto_year
##        n  missing distinct     Info     Mean      Gmd      .05
```

```
.10
##     1000        0       21    0.998     2005     6.94     1995
1997
##      .25      .50      .75      .90      .95
##     2000     2005     2010     2013     2014
##
## lowest : 1995 1996 1997 1998 1999, highest: 2011 2012 2013 2014 2015
## ---------------------------------------------------------------------
----------
```

Hemos detectado que el atributo policy_number tiene mas de 900 valores distintos y que no es un dato revelante para nuestro estudio.

Variables categoricas:

```
categoricas <- datos %>% select(-months_as_customer, -age, -
policy_number, -policy_deductable, -policy_annual_premium, -
umbrella_limit, -insured_zip, -capital.gains, -capital.loss, -
incident_hour_of_the_day, -number_of_vehicles_involved, -bodily_injuries,
-witnesses, -total_claim_amount, -injury_claim, -property_claim, -
vehicle_claim, -auto_year)
apply(categoricas,2, function(x) length(unique(x)))
```

```
##        policy_bind_date            policy_state
policy_csl
##                     951                       3
3
##             insured_sex insured_education_level
insured_occupation
##                       2                       7
14
##         insured_hobbies    insured_relationship
incident_date
##                      20                       6
60
##           incident_type          collision_type
incident_severity
##                       4                       4
4
##   authorities_contacted          incident_state
incident_city
##                       5                       7
7
##       incident_location         property_damage
police_report_available
##                    1000                       3
3
##               auto_make              auto_model
fraud_reported
##                      14                      39
2
```

Limpieza de datos

Hemos identificado que algunas columnas de cadenas tienen muchos valores distintos (900+): incident_location

Se han identificado otras variables como irrelevantes: policy_csl

Además, se existen variables que dependen unas de otras y que no aportan informacion relevante para este estudio.Auto_make y auto_model, no necesitamos saber el modelo del coche.Incident_type y collision_type, el tipo de colision solo está informado segun el tipo de incidente.

```
table(datos[,c(19,20)])
```

```
##                            collision_type
## incident_type                 ? Front Collision Rear Collision Side
Collision
##   Multi-vehicle Collision    0             115            152
152
##   Parked Car                84               0              0
0
##   Single Vehicle Collision   0             139            140
124
##   Vehicle Theft             94               0              0
0
```

Dataset de trabajo:

Eliminaremos estas columnas de nuestro conjunto de datos en el paso Procesamiento de datos para mejorar la precisión del modelo: policy_number, incident_location, policy_csl, auto_model, collision_type

Las variables que se introducen en el conjunto final son las siguientes:

```
final <- select(datos, -policy_number, -incident_location, -policy_csl, -
auto_model, -collision_type)
glimpse(final)
```

```
## Observations: 1,000
## Variables: 34
## $ months_as_customer          <int> 328, 228, 134, 256, 228, 256, 137,
165,...
## $ age                         <int> 48, 42, 29, 41, 44, 39, 34, 37,
33, 42,...
## $ policy_bind_date            <fct> 2014-10-17, 2006-06-27, 2000-09-
06, 199...
## $ policy_state                <fct> OH, IN, OH, IL, IL, OH, IN, IL,
```

```
IL, IL,...
## $ policy_deductable           <int> 1000, 2000, 2000, 2000, 1000,
1000, 100...
## $ policy_annual_premium       <dbl> 1406.91, 1197.22, 1413.14,
1415.74, 158...
## $ umbrella_limit              <int> 0, 5000000, 5000000, 6000000,
6000000, ...
## $ insured_zip                 <int> 466132, 468176, 430632, 608117,
610706,...
## $ insured_sex                 <fct> MALE, MALE, FEMALE, FEMALE, MALE,
FEMAL...
## $ insured_education_level     <fct> MD, MD, PhD, PhD, Associate, PhD,
PhD, ...
## $ insured_occupation          <fct> craft-repair, machine-op-inspct,
sales,...
## $ insured_hobbies             <fct> sleeping, reading, board-games,
board-g...
## $ insured_relationship        <fct> husband, other-relative, own-
child, unm...
## $ capital.gains               <int> 53300, 0, 35100, 48900, 66000, 0,
0, 0,...
## $ capital.loss                <int> 0, 0, 0, -62400, -46000, 0, -
77000, 0, ...
## $ incident_date               <fct> 2015-01-25, 2015-01-21, 2015-02-
22, 201...
## $ incident_type               <fct> Single Vehicle Collision, Vehicle
Theft...
## $ incident_severity           <fct> Major Damage, Minor Damage, Minor
Damag...
## $ authorities_contacted       <fct> Police, Police, Police, Police,
None, F...
## $ incident_state              <fct> SC, VA, NY, OH, NY, SC, NY, VA,
WV, NC,...
## $ incident_city               <fct> Columbus, Riverwood, Columbus,
Arlingto...
## $ incident_hour_of_the_day    <int> 5, 8, 7, 5, 20, 19, 0, 23, 21, 14,
22, ...
## $ number_of_vehicles_involved <int> 1, 1, 3, 1, 1, 3, 3, 3, 1, 1, 1,
3, 1, ...
## $ property_damage             <fct> YES, ?, NO, ?, NO, NO, ?, ?, NO,
NO, YE...
## $ bodily_injuries             <int> 1, 0, 2, 1, 0, 0, 0, 2, 1, 2, 2,
1, 1, ...
## $ witnesses                   <int> 2, 0, 3, 2, 1, 2, 0, 2, 1, 1, 2,
2, 0, ...
## $ police_report_available     <fct> YES, ?, NO, NO, NO, NO, ?, YES,
YES, ?,...
## $ total_claim_amount          <int> 71610, 5070, 34650, 63400, 6500,
64100,...
## $ injury_claim                <int> 6510, 780, 7700, 6340, 1300, 6410,
```

```
2145...
## $ property_claim              <int> 13020, 780, 3850, 6340, 650, 6410,
7150...
## $ vehicle_claim               <int> 52080, 3510, 23100, 50720, 4550,
51280,...
## $ auto_make                   <fct> Saab, Mercedes, Dodge, Chevrolet,
Accur...
## $ auto_year                   <int> 2004, 2007, 2007, 2014, 2009,
2003, 201...
## $ fraud_reported              <fct> Y, Y, N, Y, N, Y, N, N, N, N, N,
N, N, ...

write.csv(final,"insurance_claims
(insurance_claims)_insurance_claims.csv")
```