Resumen Desafío Técnico - AVLA

Fernanda Vásquez Guzmán

23 de Junio de 2019

Revisión General de la Base de Datos

Primero se hizo una revisión general de la base de datos, donde siguiendo el instructivo entregado se identificaron las variables. Para comodidad del trabajo de la base de datos, se convirtió el archivo .dms a .txt.

```
data <- read.delim("data.txt", sep = "\t")
data[1:5,1:5]
```

```
##   row_id  X0 X1  X2  X3
## 1    320 A12 30 A34 A40
## 2    211 A14 36 A32 A43
## 3     23 A12 12 A34 A41
## 4    439 A13 12 A31 A49
## 5    237 A12 21 A32 A49
```

También se redefinieron los nombres de las variables:

```
names(data)[1] <- "ID"
names(data)[2] <- "Estado.CC"
names(data)[3] <- "Duración"
names(data)[4] <- "Hist.crediticia"
names(data)[5] <- "Prop.Crédito"
names(data)[6] <- "Monto"
names(data)[7] <- "Ahorros.cuenta"
names(data)[8] <- "T.empleo.actual"
names(data)[9] <- "Tasa.%.renta"
names(data)[10] <- "Estatus.sexo"
names(data)[11] <- "Otros.deudores"
names(data)[12] <- "Tiempo.residencia"
names(data)[13] <- "Pos.nombre"
names(data)[14] <- "Edad"
names(data)[15] <- "Otros.pagos.pend"
names(data)[16] <- "Forma.vive"
names(data)[17] <- "N.cred.existentes"
names(data)[18] <- "Tipo.trabajo"
names(data)[19] <- "N.per.mantenimiento"
names(data)[20] <- "Telefono"
names(data)[21] <- "Extranjero"
names(data)[22] <- "Clasificación"
```

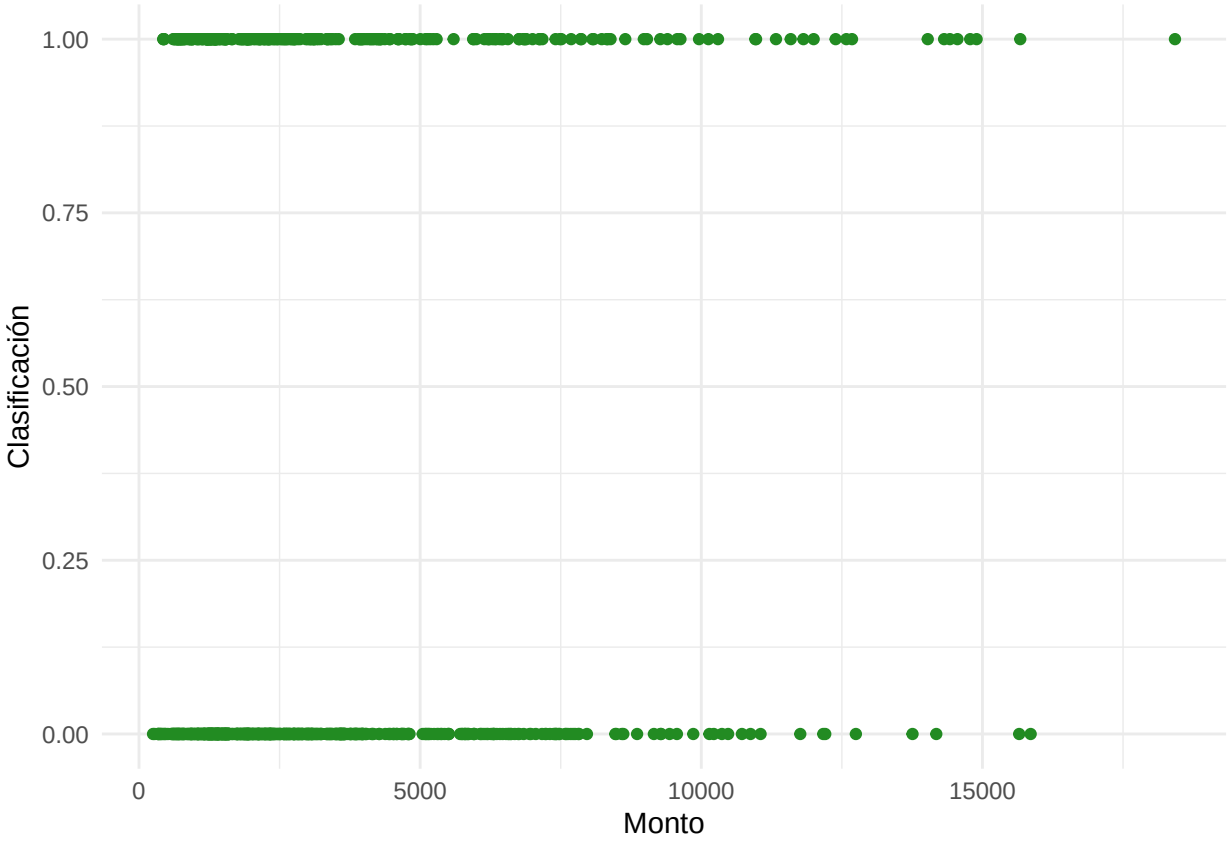
Analizando las variables, aquella que mejor puede representar el riesgo de que un cliente caiga en mora es su actual Clasificación, la cual se redefine, para que sea considerada una variable binaria. Desde este punto el cliente con buena clasificación toma el valor de 0 mientras que por el contrario el con mala clasificación toma el valor de 1.

```
data$Clasificación[data$Clasificación==1] <- 0
data$Clasificación[data$Clasificación==2] <- 1
```

También se modifica la variable Propósito del crédito que toma el valor de A410, ya que no se encuentra entre los valores que la variable puede tomar.

Análisis Descriptivo

Primero se analiza la frecuencia de esta variable dicotómica en la muestra, en relación al monto:



Donde se puede observar que no existe un punto de corte que cambie la clasificación según el monto del crédito. Además, en cuanto al grupo de personas:

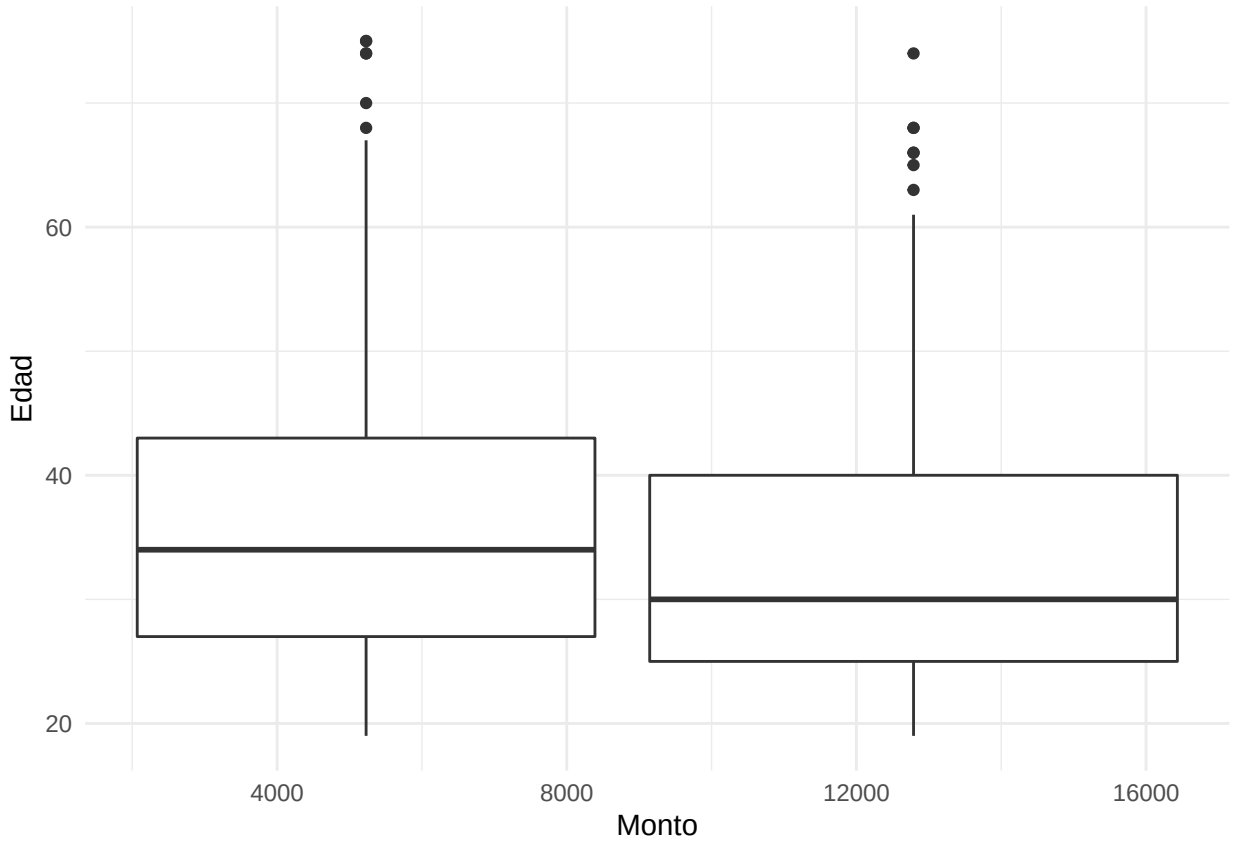
Considerando el universo de la muestra:

```
## [1] 632
```

```
## [1] 268
```

Clasificación	%
Mala	29.78%
Buena	70,2%

La mayoría de las personas tienen una buena clasificación crediticia. Y analizando las distribuciones de variables



como edad:

Se puede ver como más del 50% de la muestra se encuentra entre los 30 y 35 años para las personas con buena clasificación. También se puede ver que la frontera de edad para las personas con buena clasificación es mayor que las personas con mala clasificación, aunque esta última tiene una menor variación de edad; sin embargo, en ambos casos sobre el 75% de los datos, se encuentra una varianza mayor, ya que son menos las personas que solicitan créditos sobre los 45 años. El primer gráfico corresponde a las personas con buena calificación, mientras que el segundo a los con mala.

Modelación de los datos

Dado que se busca explicar el comportamiento de una variable dicotómica, para obtener una probabilidad teórica, se utilizarán modelos logísticos.

Dada la gran cantidad de variables, se realizara una selección de modelos, utilizando tanto un modelo logit como un modelo probit.

Para el modelo logit se obtiene:

```
##
## Call:
## glm(formula = Clasificación ~ Estado.CC + Duración + Hist.crediticia +
##     Prop.Crédito + Monto + Ahorros.cuenta + T.empleo.actual +
##     `Tasa.%.renta` + Estatus.sexo + Otros.deudores + Edad + Otros.pagos.pend +
##     Telefono + Extranjero, family = binomial(link = logit), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4009  -0.7117  -0.3684   0.7053   2.8082
```

```
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.121e+00  8.990e-01   2.360 0.018288 *
## Estado.CCA12          -4.616e-01  2.312e-01  -1.997 0.045868 *
## Estado.CCA13          -1.096e+00  3.745e-01  -2.925 0.003439 **
## Estado.CCA14          -1.790e+00  2.469e-01  -7.251 4.13e-13 ***
## Duración               3.093e-02  9.591e-03   3.225 0.001261 **
## Hist.crediticiaA31    -4.663e-01  5.898e-01  -0.791 0.429161
## Hist.crediticiaA32    -1.106e+00  4.762e-01  -2.323 0.020185 *
## Hist.crediticiaA33    -1.156e+00  5.248e-01  -2.204 0.027557 *
## Hist.crediticiaA34    -1.831e+00  4.982e-01  -3.676 0.000237 ***
## Prop.CréditoA41       -1.608e+00  3.842e-01  -4.185 2.85e-05 ***
## Prop.CréditoA42       -8.480e-01  2.695e-01  -3.147 0.001650 **
## Prop.CréditoA43       -1.078e+00  2.610e-01  -4.132 3.60e-05 ***
## Prop.CréditoA44       -1.022e+00  8.810e-01  -1.161 0.245828
## Prop.CréditoA45       -7.488e-01  6.345e-01  -1.180 0.237948
## Prop.CréditoA46        2.777e-01  4.120e-01   0.674 0.500295
## Prop.CréditoA48       -1.560e+01  4.235e+02  -0.037 0.970610
## Prop.CréditoA49       -8.188e-01  3.454e-01  -2.371 0.017756 *
## Monto                  1.253e-04  4.723e-05   2.653 0.007974 **
## Ahorros.cuentaA62     -3.522e-01  2.983e-01  -1.181 0.237647
## Ahorros.cuentaA63     -1.806e-01  3.942e-01  -0.458 0.646924
## Ahorros.cuentaA64     -1.431e+00  5.602e-01  -2.554 0.010637 *
## Ahorros.cuentaA65     -9.015e-01  2.790e-01  -3.231 0.001233 **
## T.empleo.actualA72    -4.952e-02  4.118e-01  -0.120 0.904293
## T.empleo.actualA73    -2.528e-01  3.858e-01  -0.655 0.512382
## T.empleo.actualA74    -9.209e-01  4.276e-01  -2.154 0.031276 *
## T.empleo.actualA75    -2.128e-01  3.955e-01  -0.538 0.590468
## `Tasa.%.renta`         3.379e-01  9.166e-02   3.687 0.000227 ***
## Estatus.sexoA92       -2.211e-01  3.935e-01  -0.562 0.574093
## Estatus.sexoA93       -8.540e-01  3.830e-01  -2.230 0.025750 *
## Estatus.sexoA94       -3.659e-01  4.628e-01  -0.791 0.429232
## Otros.deudoresA102     6.470e-01  4.330e-01   1.494 0.135137
## Otros.deudoresA103    -9.014e-01  4.503e-01  -2.002 0.045325 *
## Edad                  -1.556e-02  9.180e-03  -1.694 0.090187 .
## Otros.pagos.pendA142  -4.105e-01  4.429e-01  -0.927 0.353993
## Otros.pagos.pendA143  -7.706e-01  2.517e-01  -3.062 0.002202 **
## TelefonoA192          -2.907e-01  1.988e-01  -1.463 0.143525
## ExtranjeroA202        -1.531e+00  6.759e-01  -2.265 0.023484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1081.50  on 888  degrees of freedom
## Residual deviance:  791.47  on 852  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 865.47
##
## Number of Fisher Scoring iterations: 14
```

Mientras que para el modelo probit se obtiene:

```
##
```

```
## Call:
## glm(formula = Clasificación ~ Estado.CC + Duración + Hist.crediticia +
##     Prop.Crédito + Monto + Ahorros.cuenta + T.empleo.actual +
##     `Tasa.%.renta` + Estatus.sexo + Otros.deudores + Edad + Otros.pagos.pend +
##     Telefono + Extranjero, family = binomial(link = probit),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4043  -0.7281  -0.3530   0.7246   2.9265
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.241e+00  5.234e-01   2.371 0.017752 *
## Estado.CCA12         -2.844e-01  1.364e-01  -2.084 0.037126 *
## Estado.CCA13         -6.178e-01  2.160e-01  -2.860 0.004232 **
## Estado.CCA14         -1.060e+00  1.401e-01  -7.566 3.86e-14 ***
## Duración              1.730e-02  5.637e-03   3.069 0.002148 **
## Hist.crediticiaA31   -3.006e-01  3.442e-01  -0.873 0.382559
## Hist.crediticiaA32   -6.580e-01  2.766e-01  -2.379 0.017346 *
## Hist.crediticiaA33   -6.622e-01  3.053e-01  -2.169 0.030071 *
## Hist.crediticiaA34   -1.076e+00  2.882e-01  -3.735 0.000188 ***
## Prop.CréditoA41      -9.559e-01  2.178e-01  -4.388 1.14e-05 ***
## Prop.CréditoA42      -4.826e-01  1.569e-01  -3.077 0.002094 **
## Prop.CréditoA43      -6.253e-01  1.500e-01  -4.168 3.07e-05 ***
## Prop.CréditoA44      -6.219e-01  5.225e-01  -1.190 0.233944
## Prop.CréditoA45      -3.280e-01  3.631e-01  -0.903 0.366304
## Prop.CréditoA46       1.757e-01  2.395e-01   0.734 0.463039
## Prop.CréditoA48      -5.911e+00  1.055e+02  -0.056 0.955310
## Prop.CréditoA49      -4.796e-01  2.007e-01  -2.390 0.016837 *
## Monto                 7.295e-05  2.762e-05   2.641 0.008265 **
## Ahorros.cuentaA62    -2.053e-01  1.741e-01  -1.179 0.238268
## Ahorros.cuentaA63    -1.118e-01  2.231e-01  -0.501 0.616217
## Ahorros.cuentaA64    -7.633e-01  3.028e-01  -2.521 0.011705 *
## Ahorros.cuentaA65    -4.746e-01  1.574e-01  -3.015 0.002567 **
## T.empleo.actualA72   -2.422e-02  2.418e-01  -0.100 0.920200
## T.empleo.actualA73   -1.361e-01  2.267e-01  -0.600 0.548475
## T.empleo.actualA74   -5.413e-01  2.489e-01  -2.175 0.029613 *
## T.empleo.actualA75   -1.250e-01  2.317e-01  -0.539 0.589565
## `Tasa.%.renta`        1.931e-01  5.287e-02   3.653 0.000259 ***
## Estatus.sexoA92      -1.088e-01  2.326e-01  -0.468 0.639974
## Estatus.sexoA93      -4.747e-01  2.263e-01  -2.098 0.035943 *
## Estatus.sexoA94      -1.946e-01  2.727e-01  -0.714 0.475343
## Otros.deudoresA102    3.405e-01  2.562e-01   1.329 0.183797
## Otros.deudoresA103   -5.227e-01  2.565e-01  -2.038 0.041522 *
## Edad                 -8.634e-03  5.267e-03  -1.639 0.101165
## Otros.pagos.pendA142 -2.346e-01  2.607e-01  -0.900 0.368246
## Otros.pagos.pendA143 -4.607e-01  1.470e-01  -3.133 0.001730 **
## TelefonoA192         -1.729e-01  1.145e-01  -1.511 0.130912
## ExtranjeroA202       -8.794e-01  3.680e-01  -2.390 0.016855 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1081.50  on 888  degrees of freedom
## Residual deviance:  791.32  on 852  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 865.32
##
## Number of Fisher Scoring iterations: 14
```

En ambos casos las variables agregadas al modelo son las mismas, por lo que para discriminar entre ambos, se utiliza el Criterio de Información de Akaike (AIC), donde se debe escoger el modelo con el menor de los valores.

Modelo	AIC
Logit	865.47
Probit	865.32

Si bien las diferencias son mínimas, se elige el modelo Probit.

Probabilidad de caer en mora

Considerando las primeras diez observaciones de la base de datos:

```
##       Edad Monto Clasificación
##  [1,]   28  4249             2 0.720617250
##  [2,]   45  3835             1 0.061264199
##  [3,]   44  1804             1 0.044389218
##  [4,]   26   609             2 0.519634279
##  [5,]   61  2767             2 0.617719926
##  [6,]   35  8858             1 0.007869159
##  [7,]   31  7582             1 0.894258396
##  [8,]   27 14027             2 0.345776854
##  [9,]   35  1549             1 0.293053216
## [10,]   34  6614             1 0.107692903
## [11,]   22  2301             1 0.448244332
## [12,]   28  2923             1 0.769091260
## [13,]   23 15672             2 0.812906536
## [14,]   24   626             2 0.644457193
## [15,]   25  1295             2 0.120708193
## [16,]   35  3780             1 0.403417083
## [17,]   36  2247             1 0.303296340
## [18,]   36  2181             1 0.033582615
## [19,]   31  1546             1 0.151436651
## [20,]   23  1352             1 0.083174726
```

Se puede ver que las personas con una calificación crediticia mala tienen mayor probabilidad de caer en no pago.

Conclusiones variables seleccionadas

- Estado de la cuenta corriente existente: Dinero dentro de cuenta corriente disminuye probabilidad de caer en mora.

- Duración: Mayor tiempo pago de crédito, mayor probabilidad de caer en mora.
- Historia crediticia: Buena historia crediticia disminuye probabiliad de mora.
- Propósito del crédito: Probabilidad de riesgo depende del motivo del préstamo, sin embargo, invesión en activo fijo y negocios deberían disminuir la probabilidad de caer en morosidad.
- Monto: A un mayor monto de crédito, existe mayor probabilidad de caer en mora.
- Ahorros en cuenta: Mayor cantidad de ahorro disminuye probabilidad de no pago.
- Tiempo en empleo actual: Mayor tiempo en empleo actual disminuye probabilidad de no pago.
- Tasa: Mayor tasa disminuye probabilidad de pago.
- Estatus y sexo: No es altamente significativa para la determinación de la probabilidad de no pago.
- Otros deudores: Aquellos con co-deudores tienen mayor probabilidad de caer en no pago, mientras que las personas con aval disminuyen su probabilidad de no pago.
- Edad: Al aumentar la edad la probabilidad de no pago disminuye.
- Otros pagos pendientes: Tener pagos pendientes en otras entidades no aumenta la probabilidad de no pago.
- Teléfono: Las personas con teléfono propio disminuyen su probabilidad de no pago.
- Trabajador extranjero: El no ser trabajador extranjero disminuye probabilidad de no pago.