Estrategia de Construcción de Modelos para Regresión Logística: Selección Intencional

Nicolás Galindo Ramírez

2022-07-03

[ARTICULO] (doi:10.21037/atm.2016.02.15)

```
set.seed(1022409637)
df <- rnorm_multi(n = 120,
            mu = c(0.5, 300, 30, 35, 0.6),
            sd = c(0.2, 20, 5, 8, 0.15),
            r = c(0.8, 0.7, 0.5, 0.6, 0.8, 0.4, 0.3, 0.4, 0.4, 0.5),
            ## Correlación de pares de variables
            varnames = c('Compacto', 'Labranza', 'Arena', 'Arcilla','Mecanizado'))
```

```
df$Compacto <- round(df$Compacto)
df$Mecanizado <- cut(df$Mecanizado, breaks = 3,labels = c('Baja','Media','Alta'))
df
```

```
##     Compacto Labranza    Arena  Arcilla Mecanizado
## 1          0 290.3168 25.73017 35.94078      Media
## 2          0 303.4888 27.79672 27.77346       Baja
## 3          0 289.6229 27.72365 16.93676       Baja
## 4          1 342.9917 44.07114 47.70838       Alta
## 5          0 301.7775 27.21883 30.95921      Media
## 6          1 309.6803 34.70706 37.16207       Alta
## 7          1 308.9089 30.62479 32.14286       Alta
## 8          1 311.3924 34.55856 49.74534       Alta
## 9          1 316.0933 37.74212 33.64721      Media
## 10         0 272.9172 24.20817 40.67405       Alta
## 11         1 318.7808 37.62161 44.30686      Media
## 12         0 281.1278 25.94222 45.47358       Alta
## 13         1 317.0829 29.24932 38.97301      Media
## 14         1 356.1490 43.21019 46.12275       Alta
## 15         0 251.1048 18.46512 27.22439      Media
## 16         0 277.3512 28.52712 24.48133      Media
## 17         1 316.1260 34.09282 51.60994       Alta
## 18         1 286.0660 28.02977 43.83442       Alta
## 19         0 294.2069 27.75173 48.25019      Media
## 20         1 301.6528 31.47538 35.05130       Alta
## 21         1 294.5439 28.21649 37.67306       Alta
## 22         1 312.0502 32.27670 36.56516      Media
## 23         0 291.7133 28.36876 28.82534      Media
## 24         1 337.4390 43.34599 47.07522       Alta
## 25         1 327.2665 34.33753 44.76977       Alta
## 26         0 255.9013 26.97607 21.69478      Media
```

```
## 27         0 296.1913 32.52978 29.53032      Media
## 28         0 298.2821 28.98036 33.01044      Media
## 29         1 293.9268 31.46062 37.72872       Alta
## 30         0 310.6542 24.46499 33.01230      Media
## 31         1 321.6615 29.70814 30.40584       Alta
## 32         1 313.2503 30.95822 35.28499      Media
## 33         0 308.4146 34.49930 43.68971      Media
## 34         1 327.9101 36.66322 41.08379       Alta
## 35         1 317.7257 37.46228 39.85232      Media
## 36         1 301.7799 32.46582 40.04391       Alta
## 37         1 320.0277 31.81755 37.39339       Alta
## 38         0 280.1547 22.24078 24.41090      Media
## 39         0 280.8083 26.44930 23.00756      Media
## 40         0 259.5504 20.83663 13.19103      Media
## 41         1 310.2770 36.27590 33.75883      Media
## 42         0 291.2206 27.28889 25.51914       Baja
## 43         0 274.6186 24.28292 25.01714      Media
## 44         1 294.9068 36.64502 46.59897       Alta
## 45         1 323.9199 37.06746 36.84831       Alta
## 46         1 317.0736 38.46807 34.82977      Media
## 47         0 288.1622 22.18756 29.24799      Media
## 48         1 313.8991 37.89404 29.67997       Alta
## 49         0 276.3769 25.58674 35.12593      Media
## 50         0 285.0919 23.21365 29.24361      Media
## 51         0 265.1603 20.07661 26.32130       Alta
## 52         1 316.7501 33.53345 39.25440       Alta
## 53         0 285.1367 22.77083 34.62273      Media
## 54         1 335.8783 34.38577 45.65230       Alta
## 55         1 293.2915 29.65651 35.78616      Media
## 56         0 302.6336 29.07001 39.15042      Media
## 57         0 282.2286 27.74322 49.11784      Media
## 58         1 321.2227 31.00456 53.10334       Alta
## 59         1 315.4196 31.25269 53.08008       Alta
## 60         0 297.9909 25.59066 38.16370      Media
## 61         0 288.9911 31.96724 36.66260      Media
## 62         1 323.9323 35.95791 22.12337       Alta
## 63         1 305.9364 36.64046 35.59246      Media
## 64         0 268.7993 21.83665 14.48829       Baja
## 65         0 272.8860 24.12004 24.57391      Media
## 66         0 305.5650 30.74054 28.60551       Baja
## 67         1 281.9276 23.41344 38.20967       Alta
## 68         1 324.4553 35.84415 41.21267       Alta
## 69         0 304.5508 24.55070 26.49274       Baja
## 70         0 291.5669 27.61701 20.60431       Baja
## 71         1 345.5131 39.18266 43.54466       Alta
## 72         1 324.7646 36.23892 51.49996       Alta
## 73         1 331.4904 36.35074 41.81468       Alta
## 74         0 306.5237 32.06683 33.01666      Media
## 75         0 284.1244 23.65189 23.58478      Media
## 76         0 289.2180 27.33251 35.71818      Media
## 77         0 272.5110 25.86711 35.91715       Alta
## 78         1 300.6672 31.58012 33.75480      Media
## 79         1 310.8380 33.71261 48.79810       Alta
## 80         0 259.4902 24.97001 30.16777      Media
```

```
## 81         1 302.9789 33.04247 51.21253      Media
## 82         0 270.6234 21.93764 39.16714      Media
## 83         0 260.5784 24.50353 25.33571       Baja
## 84         0 259.0185 21.23153 24.21210       Baja
## 85         0 300.5805 30.44380 33.64123      Media
## 86         0 279.5347 23.68147 44.33987      Media
## 87         0 300.4289 28.91173 38.86271       Alta
## 88         1 325.6713 36.70352 41.74955       Alta
## 89         1 316.6257 32.47942 42.07620       Alta
## 90         0 267.1163 20.99090 36.70319      Media
## 91         0 283.7544 24.61160 32.17884      Media
## 92         0 288.9365 27.29195 38.64784      Media
## 93         0 277.4829 34.22339 34.56423       Alta
## 94         1 345.2133 38.12522 35.90748       Baja
## 95         1 294.2250 32.51200 55.35452      Media
## 96         0 325.0231 33.10754 39.21558      Media
## 97         0 279.9235 25.01150 32.36371       Baja
## 98         0 276.4320 27.41718 42.74673       Alta
## 99         1 311.9601 33.69218 44.05596       Alta
## 100        0 318.9084 39.00932 27.39908       Baja
## 101        0 266.0649 22.52919 29.43001      Media
## 102        1 320.5073 30.36830 20.71178      Media
## 103        0 318.7230 34.87505 48.67265       Alta
## 104        1 290.5087 26.80277 37.55585      Media
## 105        1 316.8262 32.53320 47.27949       Alta
## 106        0 291.5872 29.18598 40.57507      Media
## 107        1 316.6685 35.35891 30.44605      Media
## 108        1 284.9191 30.73893 35.69240      Media
## 109        1 299.2699 27.64507 42.76889      Media
## 110        0 281.8895 23.68858 38.86506      Media
## 111        1 306.5806 33.08293 37.15498      Media
## 112        1 317.7035 36.63353 48.29400       Alta
## 113        0 280.3368 33.07499 34.34938      Media
## 114        0 278.2350 23.14372 25.40216      Media
## 115        0 288.9289 28.30673 31.86957      Media
## 116        1 303.4105 30.01436 29.47935       Alta
## 117        1 320.7502 34.54654 35.72479       Alta
## 118        0 285.9962 26.19042 42.46509      Media
## 119        0 292.7860 32.70153 32.05938      Media
## 120        1 302.2596 34.56451 22.75500       Baja
```

Análisis univariado

```
univariable_labranza <- glm(df$Compacto ~ df$Labranza, family = binomial, data=df)
summary(univariable_labranza)
```

Labranza:

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.3963  -0.5649  -0.1183   0.5451   2.1258
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -34.61072    6.02279  -5.747 9.10e-09 ***
## df$Labranza   0.11514    0.02003   5.748 9.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  95.264  on 118  degrees of freedom
## AIC: 99.264
##
## Number of Fisher Scoring iterations: 5
```

```
univariable_arena <- glm(Compacto ~ Arena, family = binomial, data=df)
summary(univariable_arena)
```

Arena:

```
##
## Call:
## glm(formula = Compacto ~ Arena, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6787  -0.6017  -0.1756   0.6100   2.4455
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.69429    2.21436  -5.733 9.88e-09 ***
## Arena         0.41667    0.07235   5.759 8.48e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  98.861  on 118  degrees of freedom
## AIC: 102.86
##
## Number of Fisher Scoring iterations: 5
```

```
univariable_arcilla <- glm(Compacto ~ Arcilla, family = binomial, data=df)
summary(univariable_arcilla)
```

Arcilla:

```
##
## Call:
## glm(formula = Compacto ~ Arcilla, family = binomial, data = df)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8488  -0.9719  -0.3895   1.0263   2.0362
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.45266    1.03584  -4.299 1.72e-05 ***
## Arcilla      0.12138    0.02799   4.337 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.22  on 119  degrees of freedom
## Residual deviance: 141.39  on 118  degrees of freedom
## AIC: 145.39
##
## Number of Fisher Scoring iterations: 3
```

```
univariable_meca <- glm(Compacto ~ Mecanizado, family = binomial, data=df)
summary(univariable_meca)
```

Mecanizado

```
##
## Call:
## glm(formula = Compacto ~ Mecanizado, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8586  -0.8555  -0.5780   0.6257   1.9348
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.7047     0.7687  -2.218 0.026576 *
## MecanizadoMedia   0.8880     0.8166   1.087 0.276837
## MecanizadoAlta    3.2362     0.8619   3.755 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.22  on 119  degrees of freedom
## Residual deviance: 129.70  on 117  degrees of freedom
## AIC: 135.7
##
## Number of Fisher Scoring iterations: 4
```

Con respecto al analisis univariado se evidencia que todas las variables parecieran estar relacionadas con la compactación del suelo, exceptuando la labranza media, la cual pareciera no estar relacionada. Por otro lado, este analisis no es tan profundo por lo que se realiza un analisis multivariado.

#Analisis Multivariado

```
model1 <- glm(Compacto ~ Labranza + Arena + Arcilla + Mecanizado, family = binomial, data = df)
summary(model1)
```

Este paso se ajusta al modelo multivariable que comprende todas las variables identificadas en el paso anterior.

```
##
## Call:
## glm(formula = Compacto ~ Labranza + Arena + Arcilla + Mecanizado,
##     family = binomial, data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.83177  -0.46143  -0.04424   0.35134   2.00131
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -32.74578    7.50663  -4.362 1.29e-05 ***
## Labranza          0.07883    0.02755   2.861  0.00422 **
## Arena             0.19226    0.09782   1.965  0.04936 *
## Arcilla           0.02484    0.04636   0.536  0.59220
## MecanizadoMedia   1.90035    1.39344   1.364  0.17264
## MecanizadoAlta    3.69681    1.52594   2.423  0.01541 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  72.922  on 114  degrees of freedom
## AIC: 84.922
##
## Number of Fisher Scoring iterations: 6
```

```
model2 <- glm(Compacto ~ Labranza + Arena + Mecanizado, family = binomial, data = df)
summary(model2)
```

Se elimina la variable con el pvalue más alto

```
##
## Call:
## glm(formula = Compacto ~ Labranza + Arena + Mecanizado, family = binomial,
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.75863  -0.42163  -0.04717   0.35675   2.02121
```

```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -32.14105    7.29370  -4.407 1.05e-05 ***
## Labranza          0.07836    0.02730   2.870  0.00411 **
## Arena             0.19851    0.09704   2.046  0.04080 *
## MecanizadoMedia   2.15319    1.30204   1.654  0.09819 .
## MecanizadoAlta    4.02621    1.39596   2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  73.209  on 115  degrees of freedom
## AIC: 83.209
##
## Number of Fisher Scoring iterations: 6
```

Delta Beta

```
delta.coef <- abs((coef(model2)-coef(model1)[-c(5)])/coef(model1)[-c(5)])
round(delta.coef, 3)
```

```
##     (Intercept)        Labranza           Arena MecanizadoMedia  MecanizadoAlta
##           0.018           0.006           0.032          85.698           0.089
```

```
model_finalhibrido <- glm(Compacto ~ Labranza + Arena + Mecanizado, family = binomial, data = df)
summary(model_finalhibrido)
```

Se decide no eliminar ninguna variable, debido a que los cambios al eliminar la variable de pvalue más alto seria muy grande, lo cual afecta al modelo. En otras palabras el modelo se volveria inestable, por regla del 20%.

```
##
## Call:
## glm(formula = Compacto ~ Labranza + Arena + Mecanizado, family = binomial,
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.75863  -0.42163  -0.04717   0.35675   2.02121
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -32.14105    7.29370  -4.407 1.05e-05 ***
## Labranza          0.07836    0.02730   2.870  0.00411 **
## Arena             0.19851    0.09704   2.046  0.04080 *
## MecanizadoMedia   2.15319    1.30204   1.654  0.09819 .
## MecanizadoAlta    4.02621    1.39596   2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  73.209  on 115  degrees of freedom
## AIC: 83.209
##
## Number of Fisher Scoring iterations: 6
```

```
model_final <- glm(Compacto ~ Labranza + Arena, family = binomial, data = df)
summary(model_final)
```

Sin embargo se procede a verificar el anterior supuesto:

```
##
## Call:
## glm(formula = Compacto ~ Labranza + Arena, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5958  -0.5167  -0.1036   0.5710   2.4712
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -29.19777    6.30143  -4.634  3.6e-06 ***
## Labranza      0.07366    0.02420   3.044  0.00234 **
## Arena         0.23180    0.09000   2.576  0.01001 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  88.118  on 117  degrees of freedom
## AIC: 94.118
##
## Number of Fisher Scoring iterations: 6
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lrtest(model_finalhibrido, model_final)
```

```
## Likelihood ratio test
##
## Model 1: Compacto ~ Labranza + Arena + Mecanizado
## Model 2: Compacto ~ Labranza + Arena
##   #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
## 1   5 -36.605
## 2   3 -44.059 -2 14.909  0.0005789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

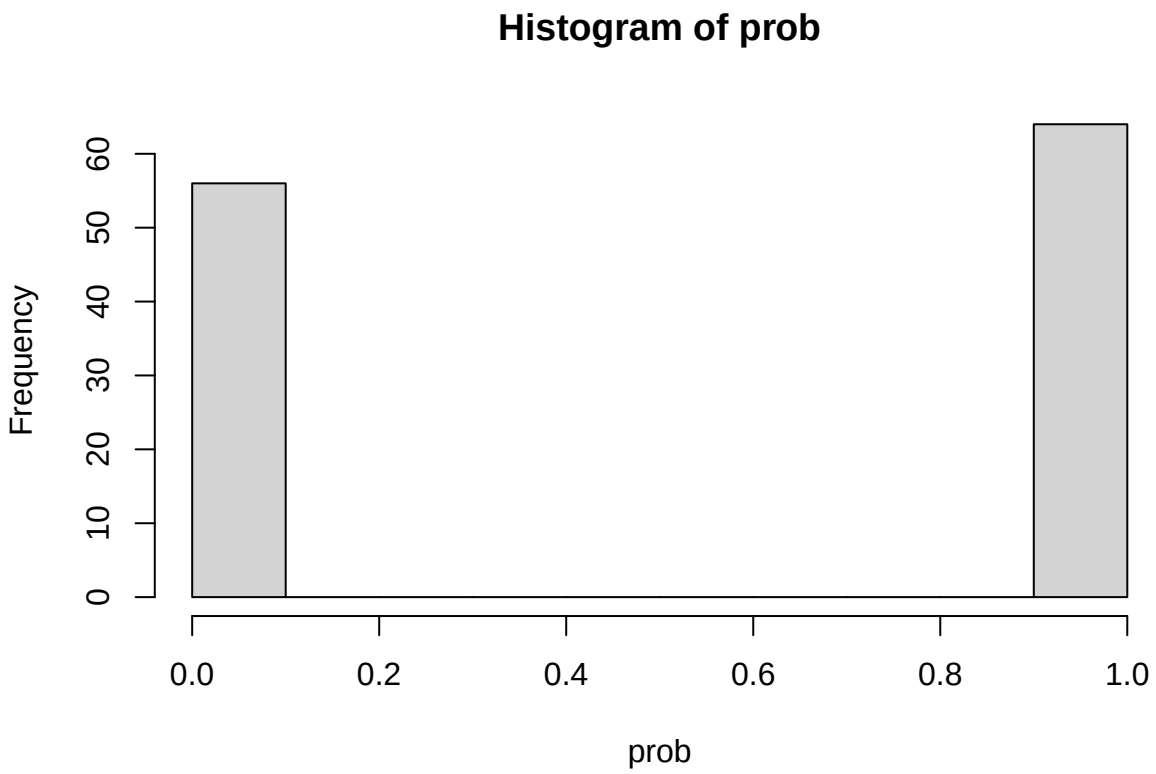
```
anova(model_finalhibrido, model_final, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Compacto ~ Labranza + Arena + Mecanizado
## Model 2: Compacto ~ Labranza + Arena
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       115     73.209
## 2       117     88.118 -2  -14.909 0.0005789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se confirma que los modelos no son estadisticamente iguales (pvalue $>$ 0,05), por lo que se rectifica que se debe trabajar con el modelo final hibrido, es decir, sin remover la variable "Mecanizado"

Supuestos de Linealidad

```
pred <- model_finalhibrido$fitted.values
prob <- ifelse(pred < 0.5, 1, 0)
hist(prob)
```



###

Utilizando las probabilidades predichas, por lo que no se tiene una probabilidad (pr) tal como se tiene en el articulo.

```
par(mfrow = c(1,3))
media_L <- mean(df$Labranza)
colores <- ifelse(df$Labranza < media_L, 'blue', 'red')
plot(model_finalhibrido$fitted.values, cex = (df$Labranza * 0.005), pch = 19, col = colores)
abline(h = 0.5, cex = 1.5, col = 'green')

media_A <- mean(df$Arena)
colores <- ifelse(df$Arena < media_A, 'blue', 'red')
plot(model_finalhibrido$fitted.values, cex = (df$Arena * 0.05), pch = 19, col = colores)
abline(h = 0.5, cex = 1.5, col = 'green')

media_M <- mean(df$Mecanizado)
```

```
## Warning in mean.default(df$Mecanizado): argument is not numeric or logical:
## returning NA
```

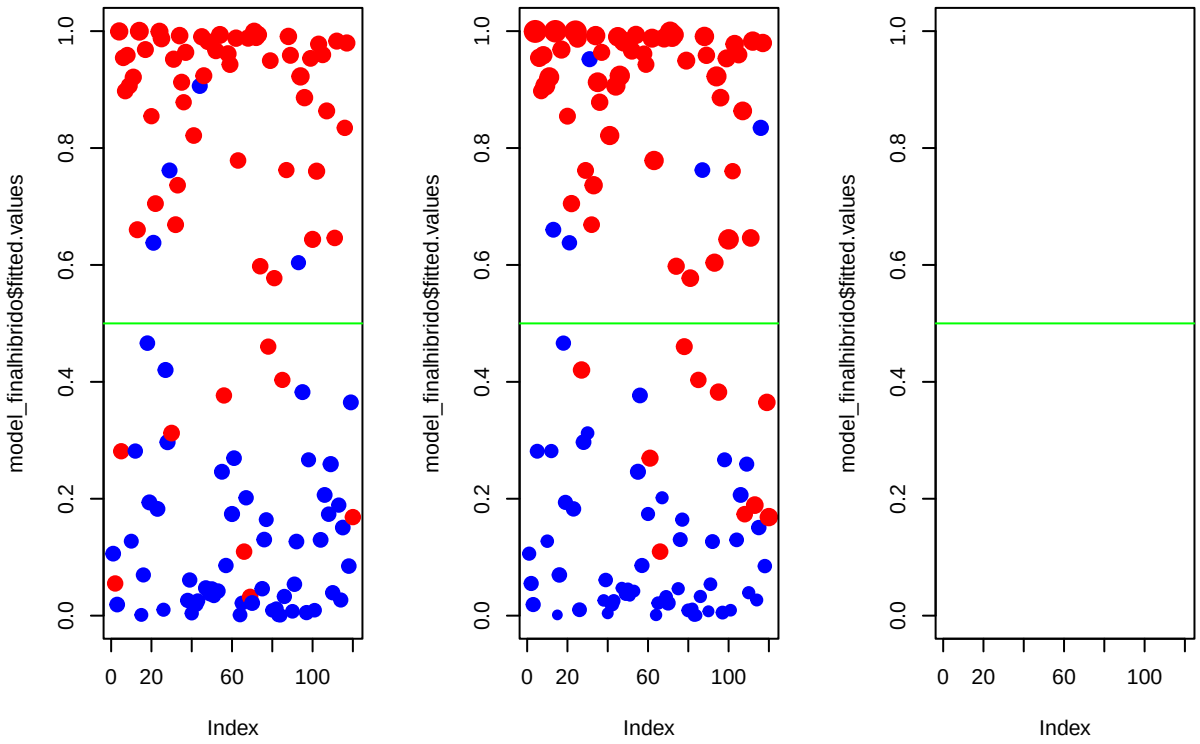
```
colores <- ifelse(df$Mecanizado < media_M, 'blue', 'red')
```

```
## Warning in Ops.factor(df$Mecanizado, media_M): '<' not meaningful for factors
```

```
plot(model_finalhibrido$fitted.values, cex = (df$Mecanizado*1000), pch = 19, col = colores)
```

```
## Warning in Ops.factor(df$Mecanizado, 1000): '*' not meaningful for factors
```

```
abline(h = 0.5, cex = 1.5, col = 'green')
```



Ninguna de las variables es apropiada para discriminar.

Interacciones del Modelo

```
model_inter1 <- glm(df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado + df$Mecanizado:df$Labranza, fa
summary(model_inter1)
```

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado +
##     df$Mecanizado:df$Labranza, family = binomial, data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.79825  -0.40945  -0.07726   0.34223   2.03991
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -22.11130   18.10921  -1.221   0.2221
## df$Labranza                      0.04574    0.06055   0.755   0.4501
## df$Arena                         0.20290    0.09770   2.077   0.0378 *
## df$MecanizadoMedia              -8.92778   20.31048  -0.440   0.6603
## df$MecanizadoAlta               -7.21869   21.06006  -0.343   0.7318
## df$Labranza:df$MecanizadoMedia   0.03568    0.06560   0.544   0.5864
## df$Labranza:df$MecanizadoAlta    0.03631    0.06854   0.530   0.5963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  72.932  on 113  degrees of freedom
## AIC: 86.932
##
## Number of Fisher Scoring iterations: 6
```

```
model_inter2 <- glm(df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado + df$Arena:df$Mecanizado, fami
summary(model_inter2)
```

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado +
##     df$Arena:df$Mecanizado, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5807  -0.3905  -0.0473   0.3553   2.1630
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -31.67762   10.04501  -3.154  0.00161 **
## df$Labranza                    0.08395    0.02773   3.028  0.00247 **
## df$Arena                       0.13350    0.22882   0.583  0.55959
## df$MecanizadoMedia            -2.37235    8.41545  -0.282  0.77802
## df$MecanizadoAlta              5.71384    8.42718   0.678  0.49776
## df$Arena:df$MecanizadoMedia    0.14349    0.25358   0.566  0.57149
## df$Arena:df$MecanizadoAlta    -0.06179    0.25368  -0.244  0.80756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  71.823  on 113  degrees of freedom
## AIC: 85.823
##
## Number of Fisher Scoring iterations: 6
```

```
model_inter3 <- glm(df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado + df$Labranza:df$Arena, family
summary(model_inter3)
```

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado +
##     df$Labranza:df$Arena, family = binomial, data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.72777  -0.40545  -0.03318   0.35215   2.03379
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -52.018856  45.622468  -1.140   0.2542
## df$Labranza            0.145553   0.154183   0.944   0.3452
## df$Arena               0.849573   1.466397   0.579   0.5623
## df$MecanizadoMedia     1.990268   1.270086   1.567   0.1171
## df$MecanizadoAlta      3.905441   1.349179   2.895   0.0038 **
## df$Labranza:df$Arena  -0.002178   0.004882  -0.446   0.6555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.22  on 119  degrees of freedom
## Residual deviance:  73.01  on 114  degrees of freedom
## AIC: 85.01
##
## Number of Fisher Scoring iterations: 6
```

```
model_inter4 <- glm(df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado + df$Labranza:df$Arena:df$Mecan
summary(model_inter4)
```

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado +
##     df$Labranza:df$Arena:df$Mecanizado, family = binomial, data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.53564  -0.35987  -0.02855   0.36732   2.16881
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -57.734664  49.379113  -1.169    0.242
## df$Labranza                             0.168498   0.158447   1.063    0.288
## df$Arena                                0.999828   1.473321   0.679    0.497
## df$MecanizadoMedia                     -0.710452   6.918880  -0.103    0.918
```

```
## df$MecanizadoAlta                          5.660583   7.122902   0.795    0.427
## df$Labranza:df$Arena:df$MecanizadoBaja    -0.002793   0.004727  -0.591    0.555
## df$Labranza:df$Arena:df$MecanizadoMedia   -0.002511   0.004923  -0.510    0.610
## df$Labranza:df$Arena:df$MecanizadoAlta    -0.002998   0.004984  -0.602    0.547
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  71.971  on 112  degrees of freedom
## AIC: 87.971
##
## Number of Fisher Scoring iterations: 6
```

```
model_inter5 <- glm(df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado + df$Labranza:df$Arena+ df$Are
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_inter5)
```

```
##
## Call:
## glm(formula = df$Compacto ~ df$Labranza + df$Arena + df$Mecanizado +
##     df$Labranza:df$Arena + df$Arena:df$Mecanizado + df$Labranza:df$Mecanizado +
##     df$Labranza:df$Arena:df$Mecanizado, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6807  -0.3649   0.0000   0.3653   2.1775
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -858.33694 2282.97103  -0.376    0.707
## df$Labranza                                2.78138    7.46035   0.373    0.709
## df$Arena                                  22.50418   60.43834   0.372    0.710
## df$MecanizadoMedia                       791.30396 2284.67020   0.346    0.729
## df$MecanizadoAlta                        790.70878 2283.88813   0.346    0.729
## df$Labranza:df$Arena                      -0.07284    0.19728  -0.369    0.712
## df$Arena:df$MecanizadoMedia              -20.98717   60.50619  -0.347    0.729
## df$Arena:df$MecanizadoAlta               -21.47130   60.46909  -0.355    0.723
## df$Labranza:df$MecanizadoMedia            -2.59003    7.46608  -0.347    0.729
## df$Labranza:df$MecanizadoAlta             -2.54310    7.46380  -0.341    0.733
## df$Labranza:df$Arena:df$MecanizadoMedia    0.06880    0.19750   0.348    0.728
## df$Labranza:df$Arena:df$MecanizadoAlta     0.06918    0.19739   0.350    0.726
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.22  on 119  degrees of freedom
## Residual deviance:  69.86  on 108  degrees of freedom
## AIC: 93.86
##
## Number of Fisher Scoring iterations: 11
```

```
rta= model_finalhibrido$fitted.values
prop_ab <- rta*100
```

```
cat_Labranza <- cut(df$Labranza, breaks = 4)
cat_Arena <- cut(df$Arena, breaks=4)
df2 <- data.frame(cat_Labranza, cat_Arena, prop_ab)

tips2 <- df2 %>%
  group_by(cat_Arena, cat_Labranza) %>%
  summarise(media_prop_compacto = mean(prop_ab))
```

Se evidencia que no hay ninguna interracción util para este caso, por lo que ninguno de estos modelos se queda. Por lo que el mejor modelo sigue siendo model_finalhibrido.

```
## `summarise()` has grouped output by 'cat_Arena'. You can override using the
## `.groups` argument.
```

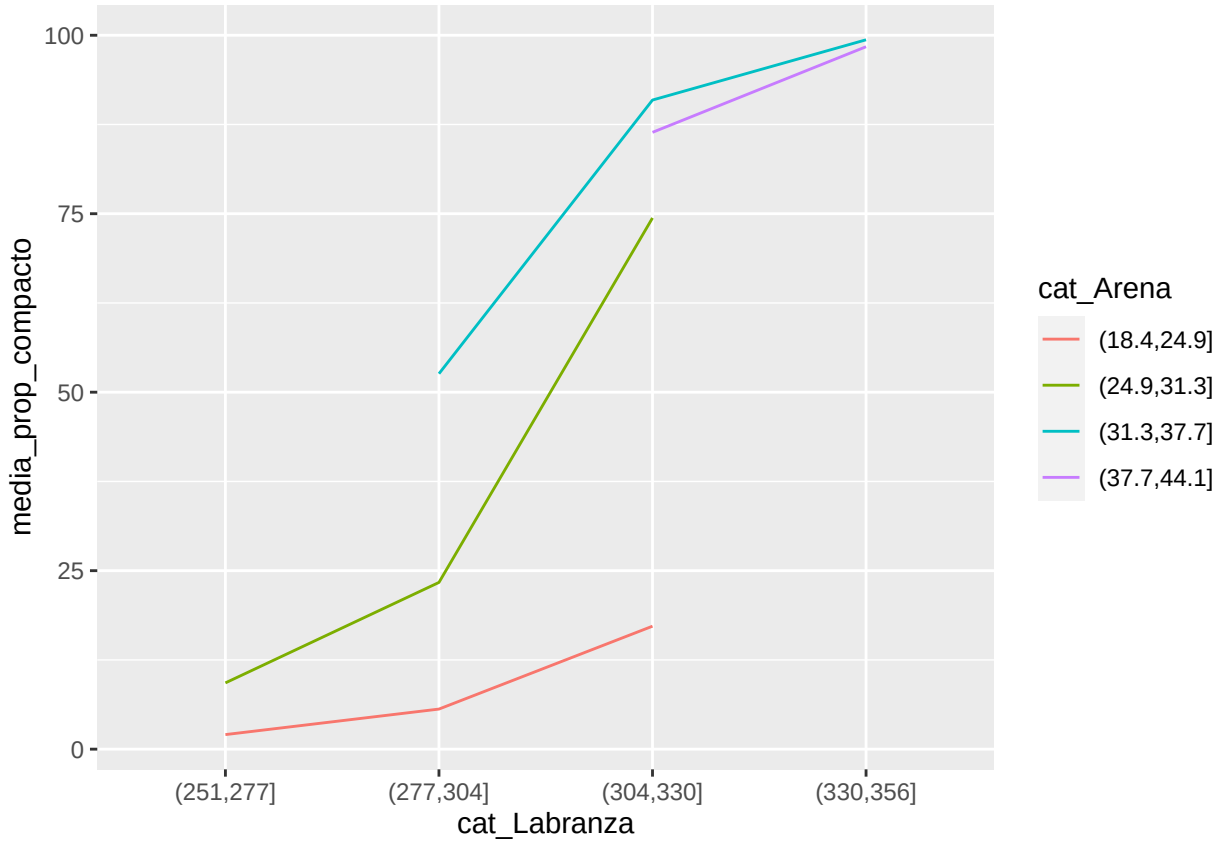
#Graficando Variables Labranza y Arena

```
library(ggplot2)
tips2$tip_groups
```

```
## Warning: Unknown or uninitialised column: `tip_groups`.
```

```
## NULL
```

```
ggplot(data = tips2) +
  aes(x = cat_Labranza, y = media_prop_compacto, color = cat_Arena) +
  geom_line(aes(group = cat_Arena))
```



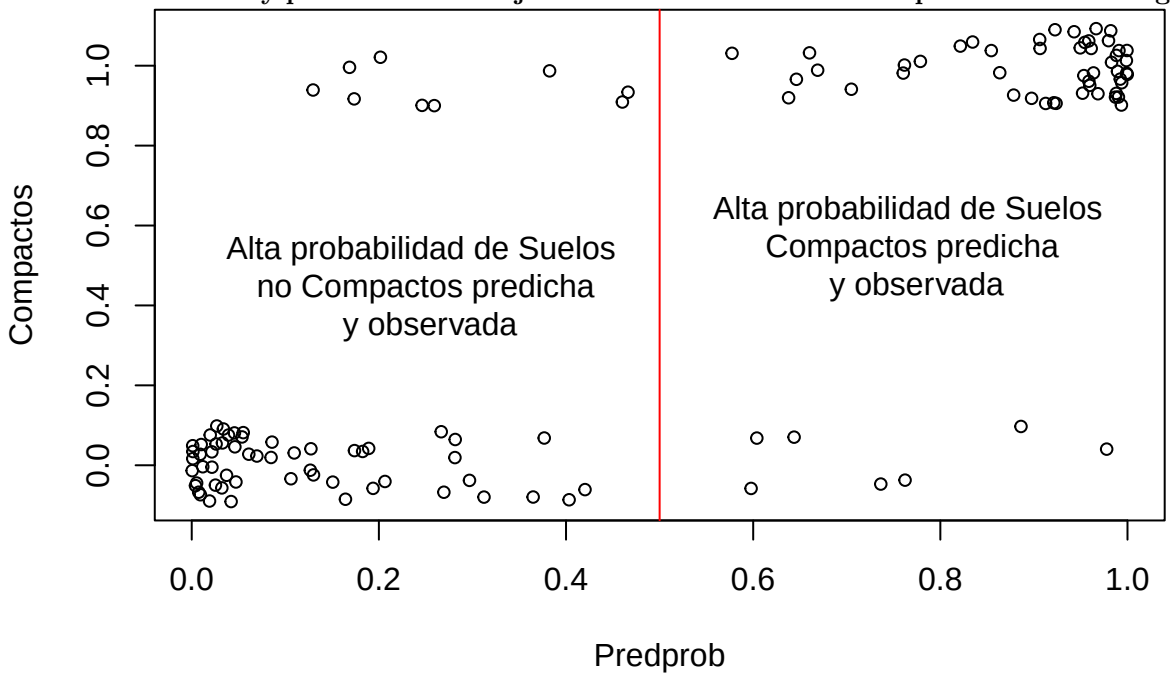
#Paso 5: Evaluación del ajuste del modelo

```
hoslem.test(model_finalhibrido$y, fitted(model_finalhibrido))
```

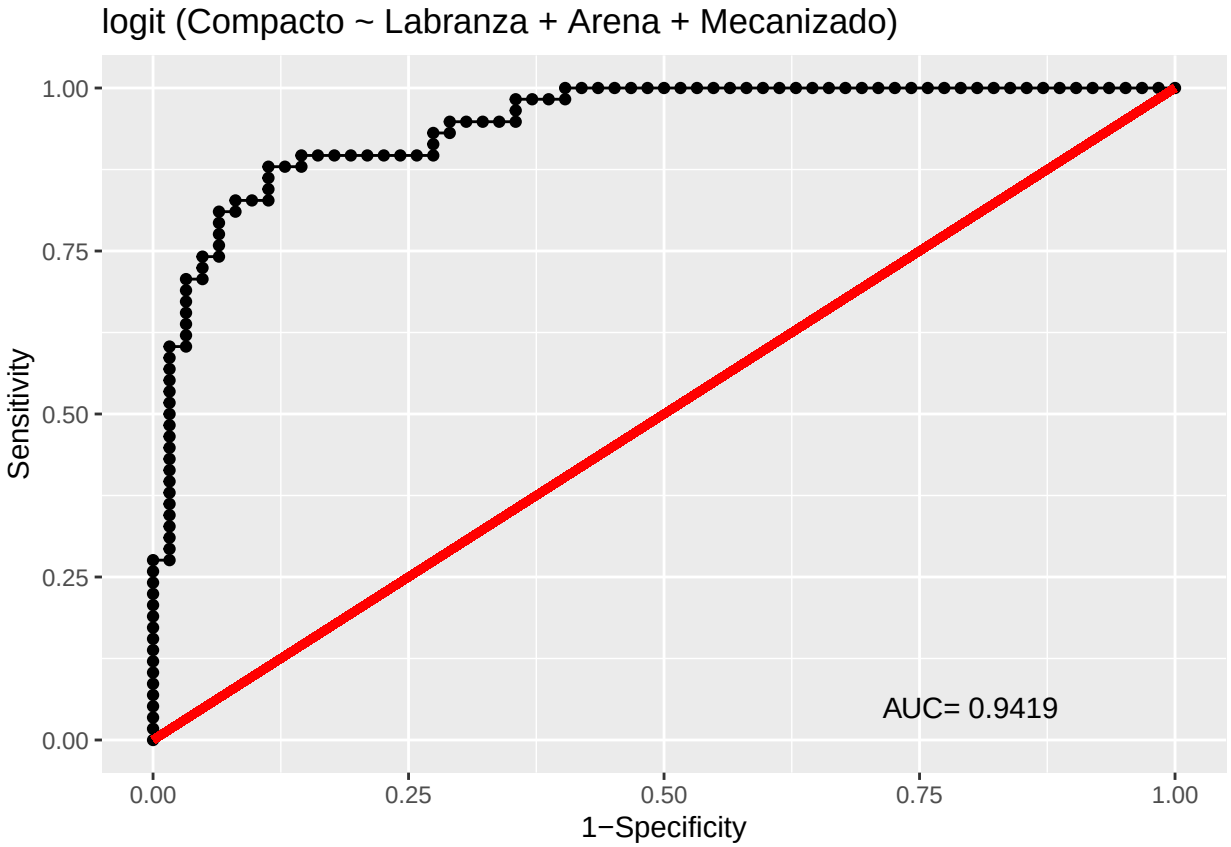
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_finalhibrido$y, fitted(model_finalhibrido)
## X-squared = 6.4323, df = 8, p-value = 0.5989
```

```
Predprob<-predict(model_finalhibrido,type="response")
plot(Predprob,jitter(as.numeric(df$Compacto), 0.5), cex=0.8,     ylab="Compactos")
abline(v = 0.5, col = 'red')
text(x = 0.77, y = 0.55, 'Alta probabilidad de Suelos \n Compactos predicha \n y observada')
text(x = 0.25, y = 0.45, 'Alta probabilidad de Suelos \n no Compactos predicha \n y observada')
```

El valor de P es 0,59, lo que nos indica que no hay una diferencia significativa entre los valores observados y predichos. Este ajuste en el modelo tambien se puede analizar en graficas




```
rocplot(model_finalhibrido)
```



En estos graficos se observa que hubo bastantes aciertos de forma predicha y observada con respecto a si el suelo esta compacto o no.