Detección de fraudes - Fundamentos de Estadística

Abraham Nieto 51556 y Alejandro Hernández 87806

Diciembre 2018

Índice

Introducción

La detección de fraude es uno de los eventos más difíciles de detectar debido a que los fraudes son eventos de baja densidad; es decir, si planteamos el evento de fraude de forma binaria, 0 (no es fraude) y 1 (sí es fraude), la proporción de 1's con respecto a los 0's es sigificativamente menor. Otra forma de representarlo es que tenemos un problema de clases no balanceadas lo cual hace difícil encontrar el evento de interés que en este caso es el fraude.

En los datos que analizaremos en el presente trabajo, la variable respuesta binaria, fraudRisk tiene media 0.0658, lo que significa que la densidad de fraudes en la base es casi de un 6.6%, lo que significa que para encontrar un caso de fraude tendríamos que revisar al menos 15 registros, esto suponiendo que los revisamos al azar.

Entonces el objetivo es calcular la propensión de que a cada cliente le hayan hecho fraude, para esto debemos comenzar analizando las distintas variables y encontrar aquellas que nos puedan servir para discriminar los casos de fraude.

Posteriormente, generaremos diversos modelos de clasificación y determinaremos la métrica adecuada para la selección del modelo óptimo.

Análisis Exploratorio

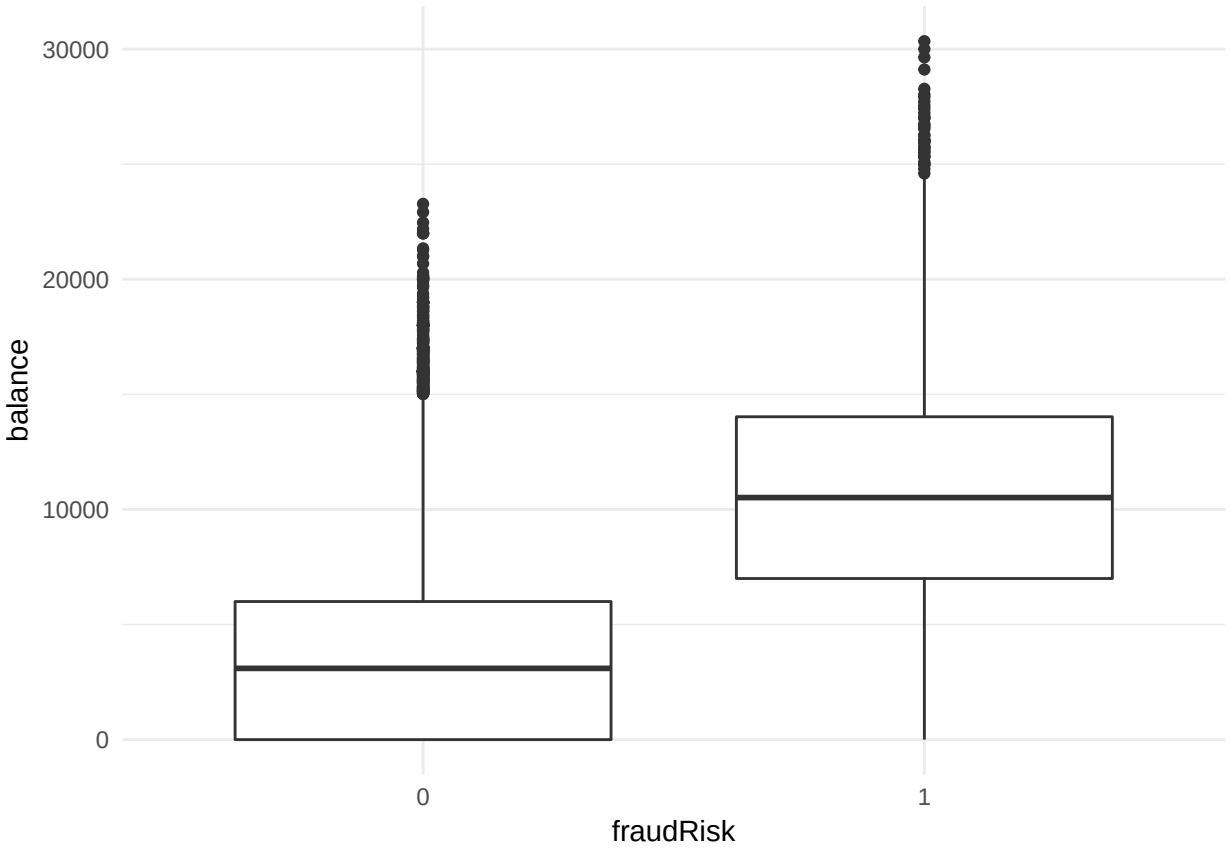
A continuación presentamos un "summary" de los datos:

```
##      gender          state          cardholder      balance
##  Min.   :1.000   Min.   : 1.00   Min.   :1.00   Min.   :    0
##  1st Qu.:1.000   1st Qu.:10.00   1st Qu.:1.00   1st Qu.:    0
##  Median :1.000   Median :24.00   Median :1.00   Median : 3683
##  Mean   :1.385   Mean   :24.75   Mean   :1.03   Mean   : 4099
##  3rd Qu.:2.000   3rd Qu.:38.00   3rd Qu.:1.00   3rd Qu.: 6000
##  Max.   :2.000   Max.   :51.00   Max.   :2.00   Max.   :30344
##     numTrans       numIntlTrans      creditLine       fraudRisk
##  Min.   :  0.00   Min.   : 0.000   Min.   : 1.000   Min.   :0.00000
##  1st Qu.: 10.00   1st Qu.: 0.000   1st Qu.: 4.000   1st Qu.:0.00000
##  Median : 20.00   Median : 0.000   Median : 6.000   Median :0.00000
##  Mean   : 29.07   Mean   : 4.083   Mean   : 9.151   Mean   :0.06058
##  3rd Qu.: 39.00   3rd Qu.: 4.000   3rd Qu.:11.000   3rd Qu.:0.00000
##  Max.   :100.00   Max.   :60.000   Max.   :75.000   Max.   :1.00000
```

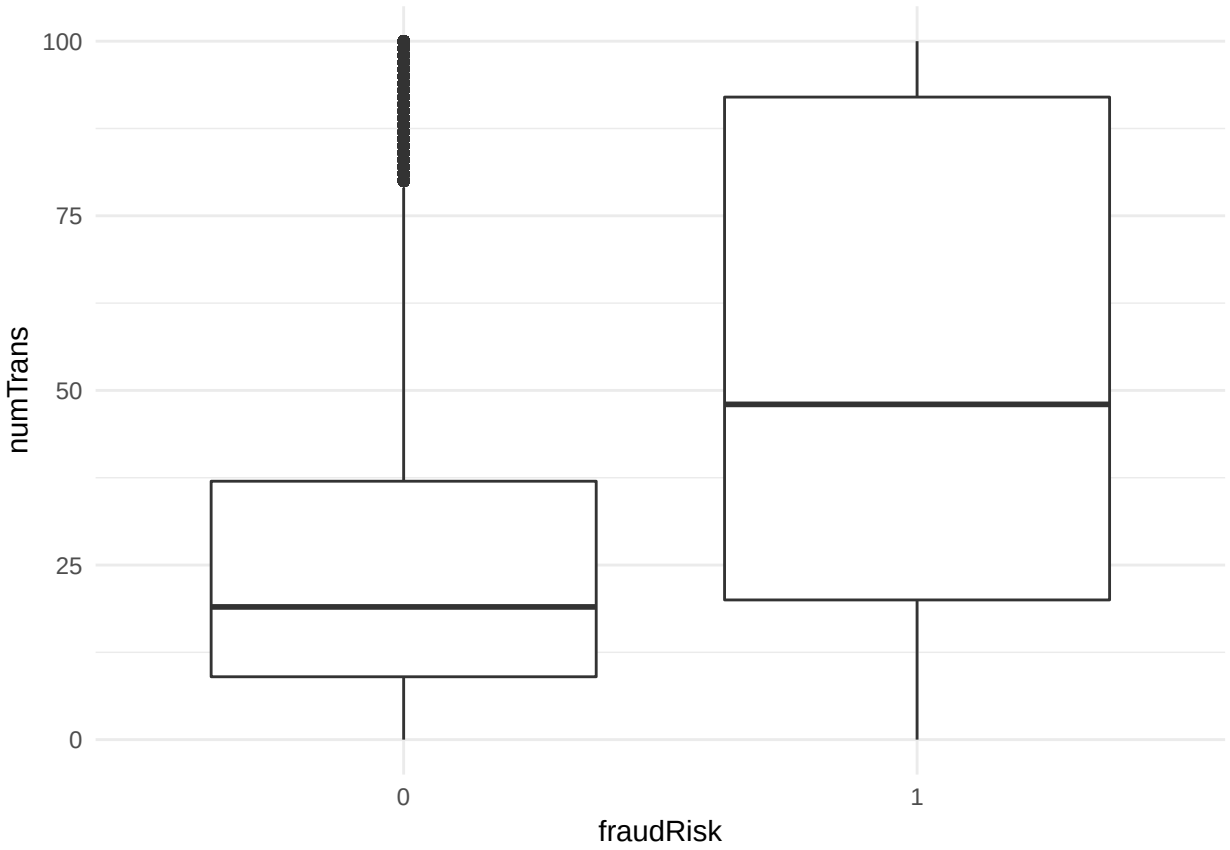
Se cuenta con información del género (1 para hombre y 2 para mujer), el estado de EUA donde reside el cliente (1 a 54), cardholder (1 titular y 2 adicional), balance de la cuenta (0 a 30344), número de transacciones nacionales (0 a 100), número de transacciones internacionales (0 a 60), línea de crédito (1 a 75) y riesgo de fraude (1 sí y 0 no).

Primero realizamos un mapa de la información para detectar si existen valores ausentes en la base que debamos imputar, lo cual no es necesario ya que notamos que no existen valores ausentes.

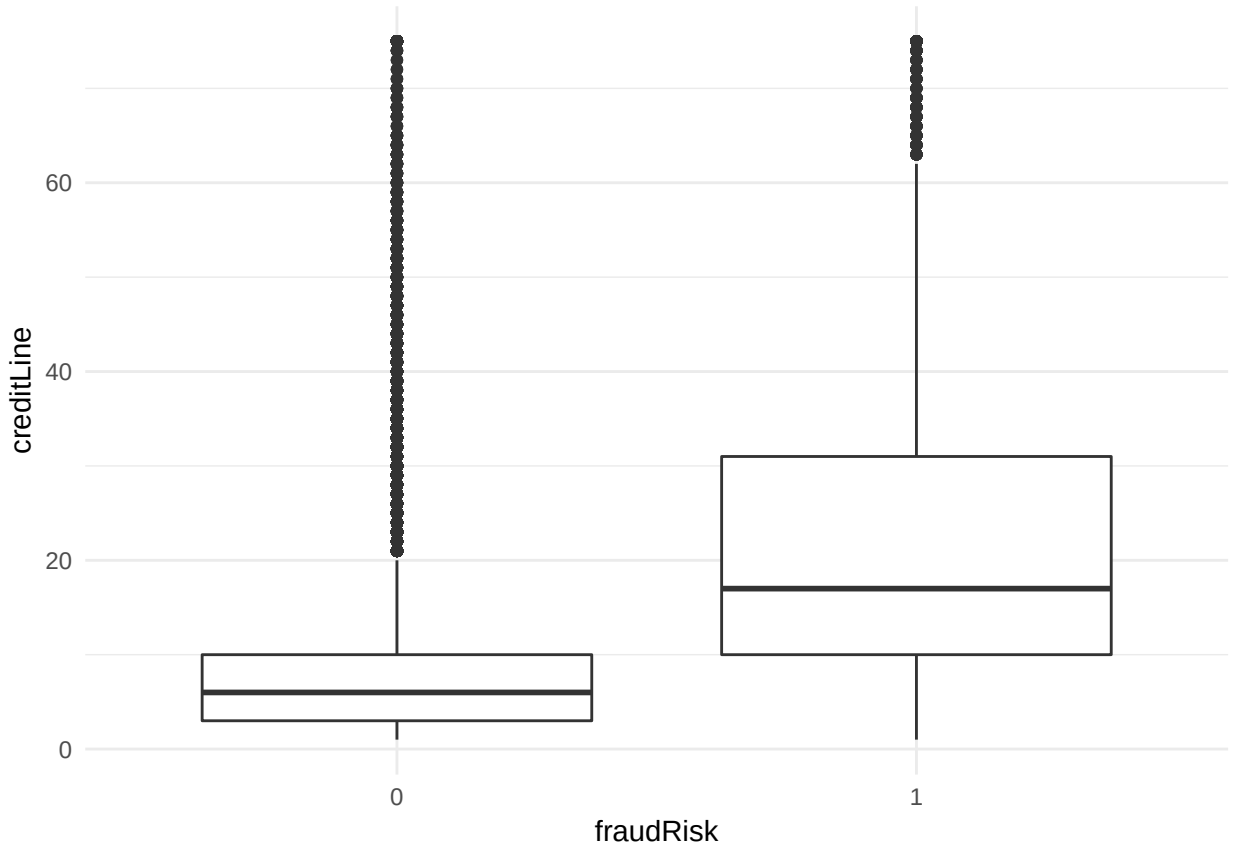
Revisando la variable balance contra la variable respuesta fraudRisk notamos que los casos de fraude se dan el 75% de las veces donde los saldos de los clientes son mayores a 7500 USD aproximadamente, mientras que en caso negativo este saldo se encuentra en la cuarta parte de los clientes. Con lo anterior, resulta claro que esta variable funciona para contrastar los casos.



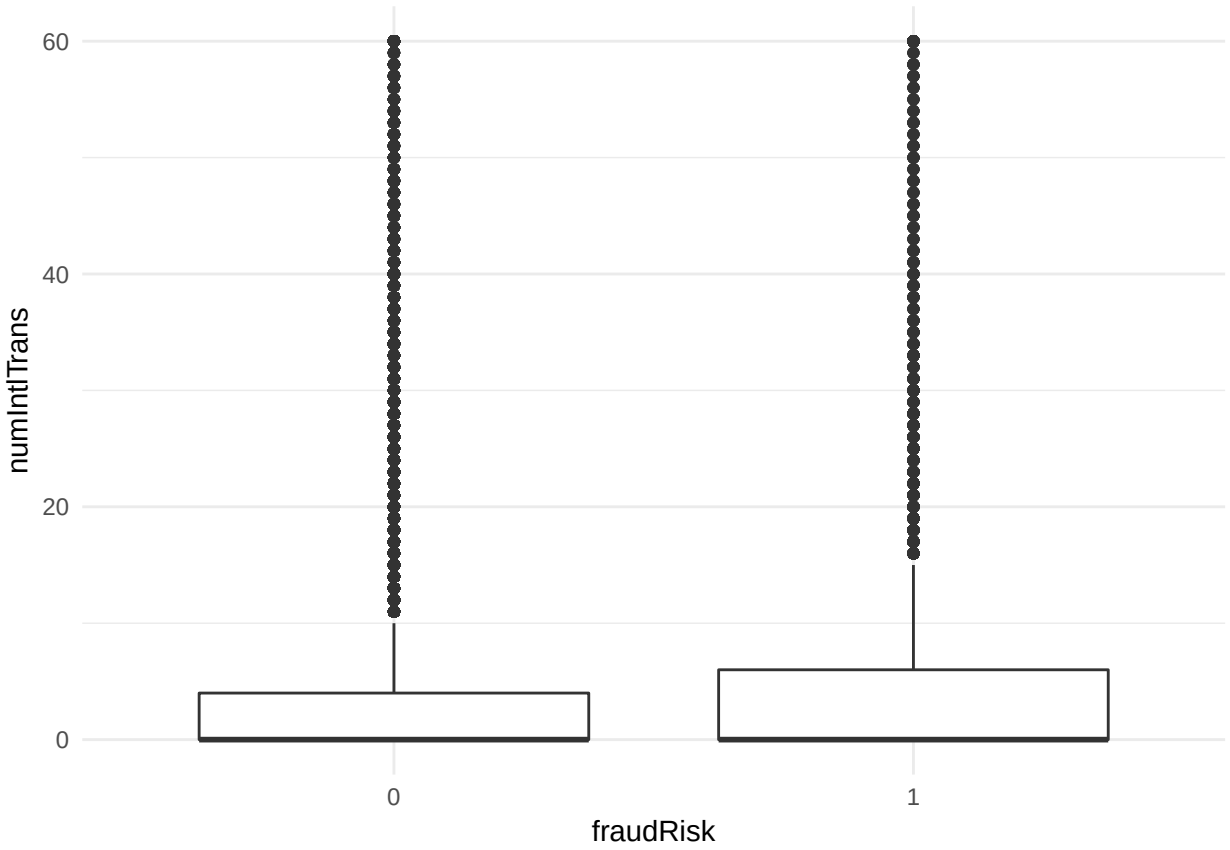
Por otro lado, analizando la variable numTrans (Número de transacciones domésticas en un periodo dado) en contraste con la variable respuesta fraudRisk apreciamos que los casos de fraude se dan el 50% de las veces donde el cliente tiene más de 50 transacciones y en los casos negativos este número de transacciones aparece en menos del 75% de los casos.



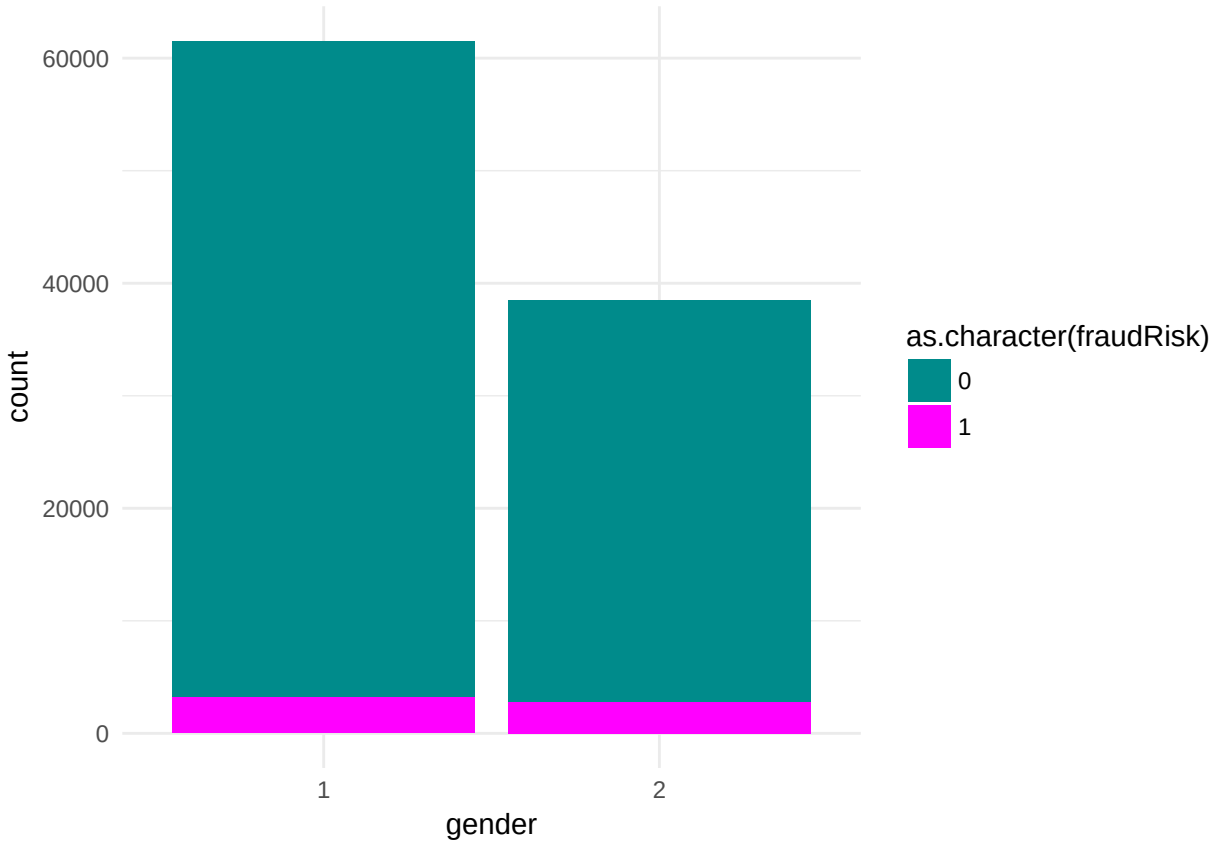
Para la variable creditLine en contraste con la variable respuesta fraudRisk, apreciamos que los casos de fraude se dan el 50% de las veces donde el cliente tiene aproximadamente una línea de crédito de 20 y en los casos negativos el 75% de los casos tiene una linea de credito de 10.



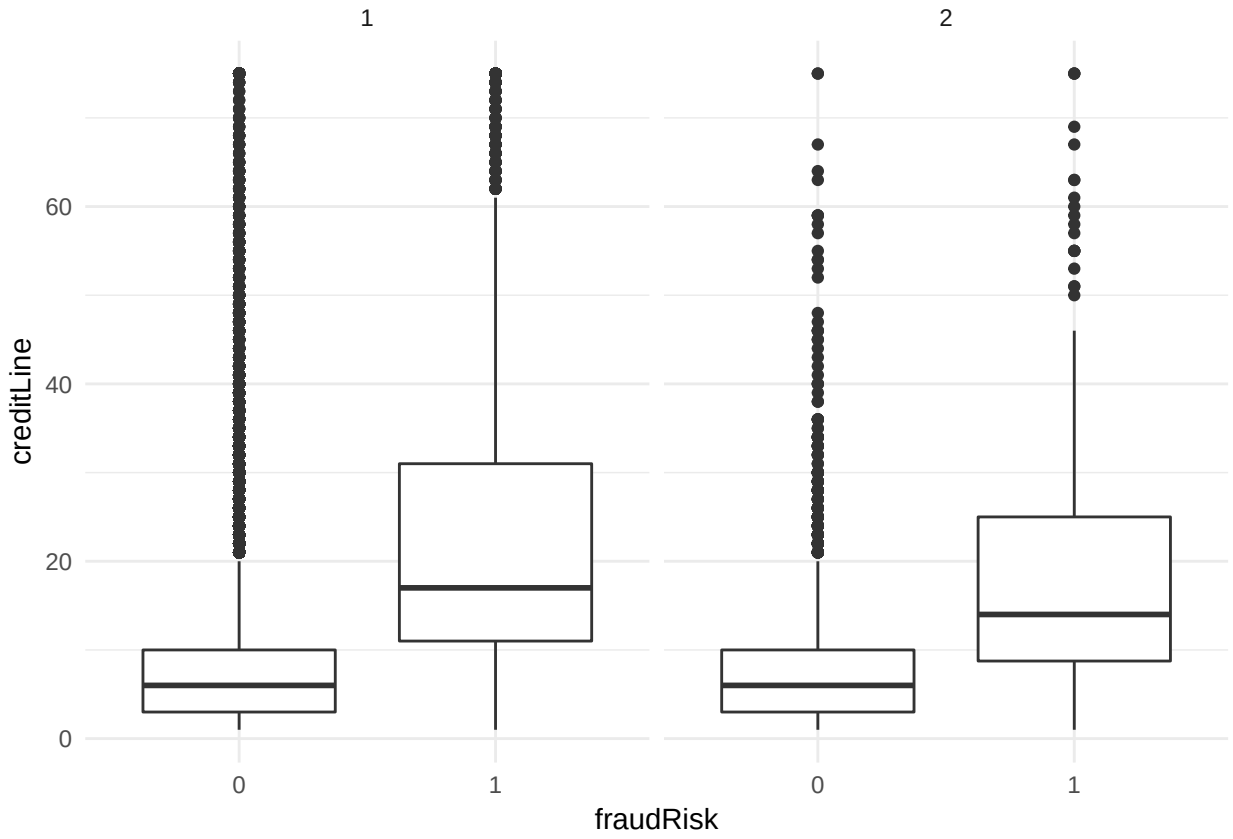
Por su parte, la variable numIntlTrans en contraste con la variable respuesta fraudRisk vemos que la distribución es casi igual lo cual nos dice que la variable no sirve para discriminar las clases de fraude.



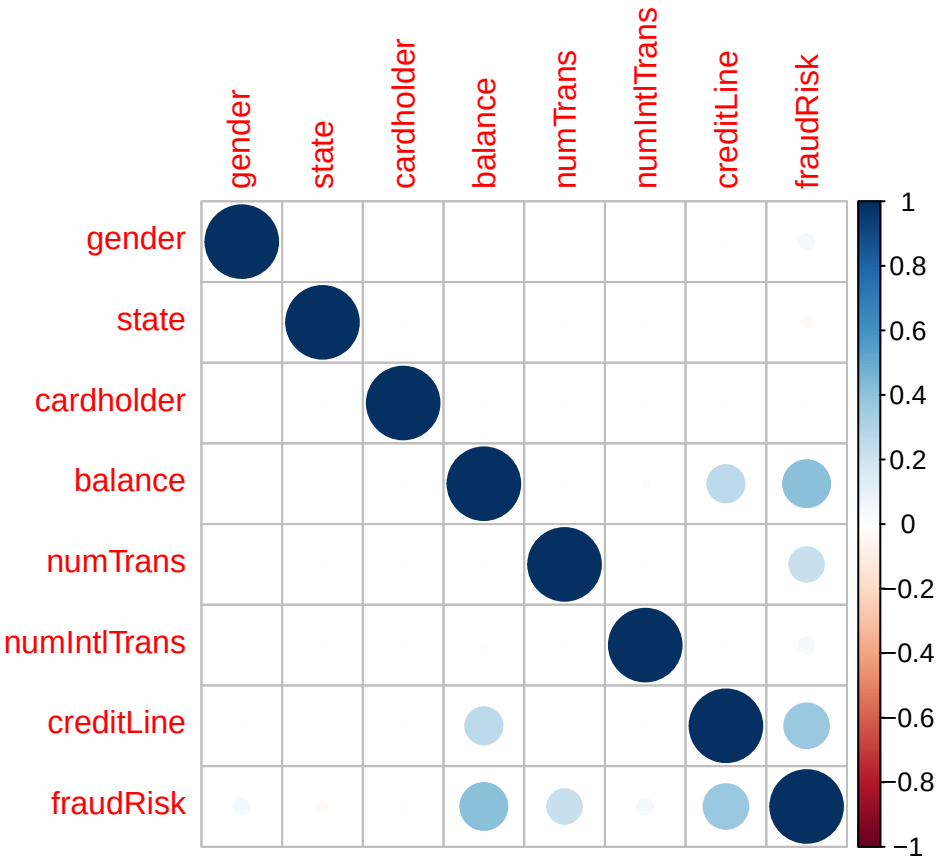
Notamos que el riesgo de fraude es bastante más alto en mujeres que en hombres, pues a pesar de que el número de mujeres es menor, el numero de fraudes que comente es similar al de los hombres.



La distinción entre los posibles valores de la variable fraudRisk se aprecia de manera más adecuada cuando se trata del titular de la tarjeta.



Notamos que las correlaciones más altas con respecto a la variable respuesta se dan con las variables balance, creditLine y numTrans.



Segmentos

De acuerdo con los comentarios del profesor vamos a segmentar con base en las variables balance y tipo de línea de crédito.

Primero segmentamos la variable balance como sigue:

balance_q	segmento	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
1	[0,2822)	40000	1.05	419	6.92	40.00
2	[2822, 4394)	20002	1.65	330	5.45	20.00
3	[4394, 7001)	21519	4.01	862	14.23	21.52
4	[7001,30344]	18479	24.07	4447	73.41	18.48

Después de segmentar la variable balance en cuatro grupos podemos observar con la ayuda de la tabla anterior que el 73% del fraude se concentra en el segmento 4 que contiene tan solo el 18% de los clientes, recordar que el segmento 4 son aquellos clientes con un balance mayor a los 7 mil dólares, también es importante notar que la tasa de fraude en este segmento es de más del 24%, lo que nos indica que es 4 veces mayor a la densidad original que es del 6%.

Ahora segmentamos la variable creditLine como sigue:

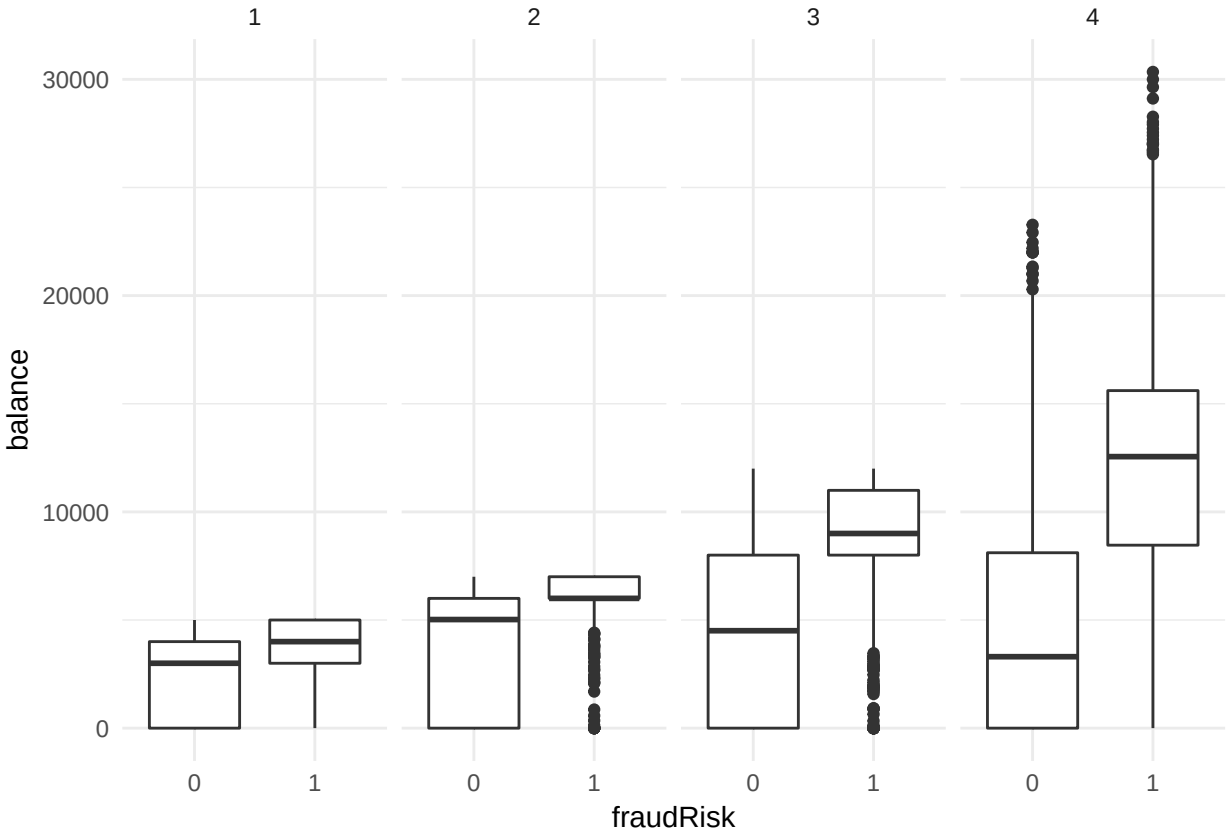
credl_q	segmento	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
1	[1, 5)	36092	0.82	296	4.89	36.09
2	[5, 7)	17670	2.05	363	5.99	17.67

credl_q	segmento	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
3	[7,12)	23296	4.76	1110	18.32	23.30
4	[12,75]	22942	18.69	4289	70.80	22.94

Haciendo 4 segmentos en los tipos de líneas de créditos podemos observar que el segemento 4 que se refiere del tipo 12 en adelante contiene el 70% de los clientes fraudulentos y representa casi el 23% de los clientes totales, además la tasa de fraude en este segmento es del 18.7% lo cual es más de 3 veces mayor que la densidad original de de 6% entonces es 3 veces más probable que haya clientes o casos de fraude en este segmento, por otro lado dentro de la segementación de esta varible el segmento 3 que representa los tipos entre 6 y 11 tiene una tasa de fraude de 4.76% la cual es menor que la densidad original al igual que los segmentos 1 y 2 con tasas de fraude de .82% y 2.05% respectivamente.

Ahora dado que nuestro objetivo es construir un modelo de propensión de fraude mezclamos los segmentos de ambas variables de tal forma que para tener mayor asertividad en la propensión se tenga un modelo para cada segmento utilizando el resto de las variables como explicativas.

Cruzando los distintos segmentos tenemos los siguientes resultados:



Si observamos los diagramas de caja y brazos de los 4 segmentos de los tipos de líneas de crédito a través de su balance o saldo vemos que para el segmento 4 la diferencia entre las distribuciones de balance de los casos de fraude y no fraude se diferencían de forma muy clara, lo cual hace pensar que en este segmento sería más sencillo encontrar un modelo que discrimine a los cientes más propensos a ser casos de fraude, en los primeros 2 segmentos pudiera ser más complejo dado que las cajas se traslapan.

Análizando los segmentos mezclados de las variables balance y creditLine tenemos lo siguiente:

balance_q	credl_q	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
4	4	9136	38.43	3511	57.96	9.14

balance_q	credl_q	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
4	3	9343	10.02	936	15.45	9.34
3	4	2916	11.21	327	5.40	2.92
3	2	9520	3.37	321	5.30	9.52
1	4	9075	3.28	298	4.92	9.07
2	4	1815	8.43	153	2.53	1.82
2	1	15034	0.89	134	2.21	15.03
3	1	6184	1.86	115	1.90	6.18
3	3	2899	3.41	99	1.63	2.90
1	3	9255	0.55	51	0.84	9.25
1	1	14874	0.32	47	0.78	14.87
2	3	1799	1.33	24	0.40	1.80
1	2	6796	0.34	23	0.38	6.80
2	2	1354	1.40	19	0.31	1.35

El segmento (4,4) (balance,creditLine) contiene el 58% de los casos de fraude mientras que su densidad de fraude es de 38.43%, este segmento es donde de manera más sencilla se pueden encontrar los casos fruadulentos, luego los segmentos (4,3) y (3,4) representan casi el 21% del total de los casos de fraude con una tasa de casos de fraude de más del 10% es decir con estos 3 segmentos se cubre casi el 79% de los casos fraudulentos, lo que estamos haciendo con estos cruces de segmentos es crear categorías de mayor a menor "facilidad" para detectar los casos, entonces la idea es que los modelos que construyamos encuentren probabilidades más asertivas en estos 2 segmentos y obviamente mucho menores en el resto de ellos. En términos de negocio queremos encontrar 3 tipos de segmentos, digamos alto, medio y bajo donde podamos saber que la probabilidad de encontrar casos de fraude va de mayor a menor, de tal forma que podamos atacar este problema de detección en cada uno de ellos.

Por tanto, se definen los siguientes segmentos:

- Alto= {segmento (4,4)}
- Medio={segmento (4,3), segmento (3,4)}
- Bajo={todos los segmentos}-{Alto,Medio}

Segmento	ctes	tasa_f	total_f	proporcion_f	proporcion_ctes
Alto	9136	38.43	3511	57.96	9.14
Bajo	78605	1.63	1284	21.20	78.61
Medio	12259	10.30	1263	20.85	12.26

Finalmente con estos 3 segmentos podemos observar que para el segmento Alto se detecta el 58% delos casos de fraude analizando sólo el 9% de los clientes, para el Medio se detecta el 21% de los fraudes con el 12.2% de los clientes ambos con densidades o tasas de fraude mayores a la original, para el segmento más bajo tiene una densidad del 1.6%.

Modelos jerárquicos

En primera instancia, cabe destacar que convertimos a dummys las variables "gender"" y "cardholder".

El Segmento lo definimos como Alto=1, Medio=2 y Bajo=3.

Asimismo, no se contempló la variable "state" ya que al considerarla en el problema jerárquico hacia que fuera muy complejo su cómputo.

Modelo beta-bernoulli

El modelo se especificó como sigue:

$$Y_i \sim Ber(p_{ij})$$

$$p_j \sim Beta(a, b)$$
$$a, b \sim gama(0.01, 0.01)$$

Clasificación global:

	0	1
0	0.9394703	0.625
1	0.0605297	0.375

Clasificación por segmento:

Alto:

	0	1
0	0.6156798	0.625
1	0.3843202	0.375

Medio:

	0
0	0.8969737
1	0.1030263

Bajo:

	0
0	0.9836652
1	0.0163348

Notamos que este modelo da resultados muy malos, pues en términos generales la sensibilidad es del 33%. Por su parte, para los segmentos medio y bajo no está clasificando ningún fraude, lo cual no es apropiado. Por lo anterior, no entraremos al análisis de los coeficientes y descartaremos este modelo.

Modelo lineal generalizado con efectos independientes

El modelo de efectos constantes se especificó de la siguiente forma:

$$Y_i \sim Ber(p_i)$$

$$logit(p_i) = \alpha_{ij} + \beta_{1j} genero_{ij} + \beta_{2j} cardholder_{ij} + \beta_{3j} balance_{ij} + \beta_{4j} numTrans_{ij} + \beta_{5j} numIntTrans_{ij} + \beta_{6j} creditline_{ij}$$

Donde:

$$\alpha_j \sim N(0, 0.001)$$

$$\beta_j \sim N(0, 0.001)$$

con $j = 1, 2, 3$ los segmentos Alto=1, Medio=2 y Bajo=3.

Revisamos que todas las cadenas se estabilizaran y que no hubiera problemas de autocorrelación.

A continuación realizamos una tabla para ver que tan bien clasificó los fraudes nuestro modelo:

	0	1
0	0.9659628	0.2530218
1	0.0340372	0.7469782

Notamos que la sensibilidad no es de lo más adecuada, pues es del 75%.

Ahora veremos el efecto por cada segmento:

	0	1
0	0.8018776	0.2268425
1	0.1981224	0.7731575

La tabla anterior nos muestra la clasificación para el segmento alto.

Ahora veamos el segmento medio:

	0	1
0	0.9244361	0.3633333
1	0.0755639	0.6366667

Finalmente, para el segmento bajo:

	0	1
0	0.9850586	0.3212121
1	0.0149414	0.6787879

El resultado es el esperado, pues el segmento alto es el que tiene la mejor sensibilidad debido a que es el segmento con mayor número de fraudes, por lo que la clasificación dentro de dicho segmento es más adecuada.

Modelo lineal generalizado con efectos intercambiables

El modelo lo especificamos como sigue:

$$Y_i \sim Ber(p_i)$$

$$logit(p_i) = \alpha_{ij} + \beta_{1j} genero_{ij} + \beta_{2j} cardholder_{ij} + \beta_{3j} balance_{ij} + \beta_{4j} numTrans_{ij} + \beta_{5j} numIntTrans_{ij} + \beta_{6j} creditline_{ij}$$

Donde:

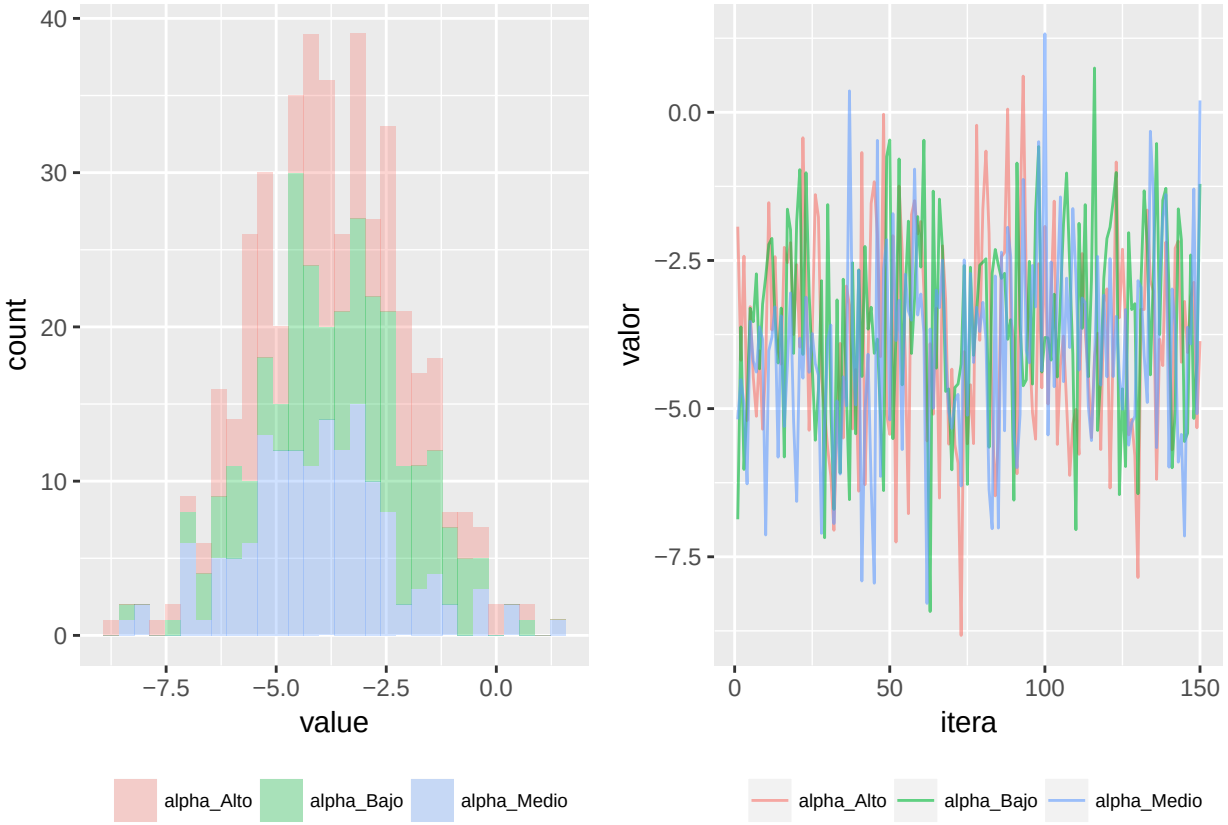
$$\alpha_j \sim N(0, \tau)$$

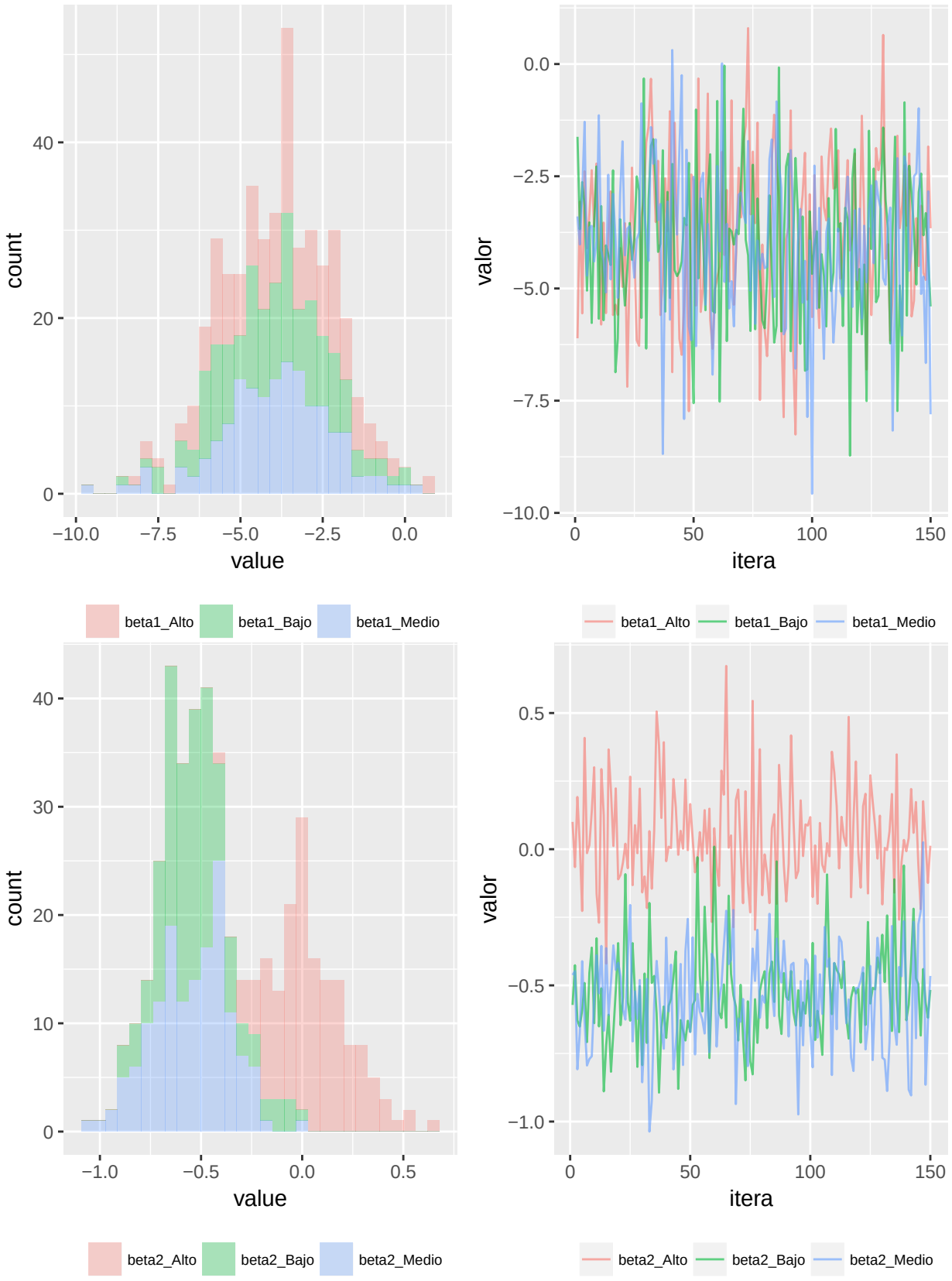
$$\beta_j \sim N(0, \tau)$$

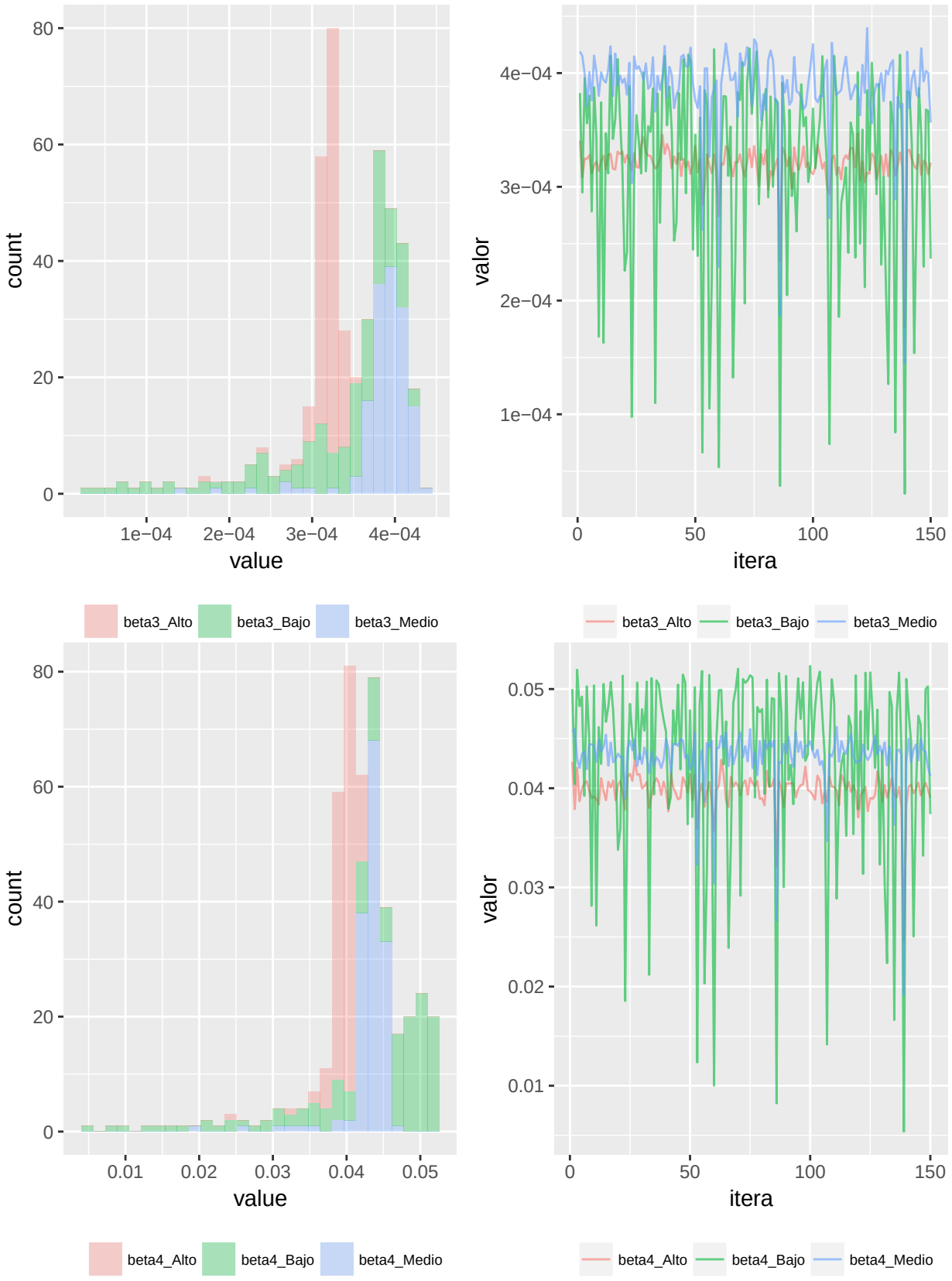
$$\tau \sim Gama(0.001, 0.001)$$

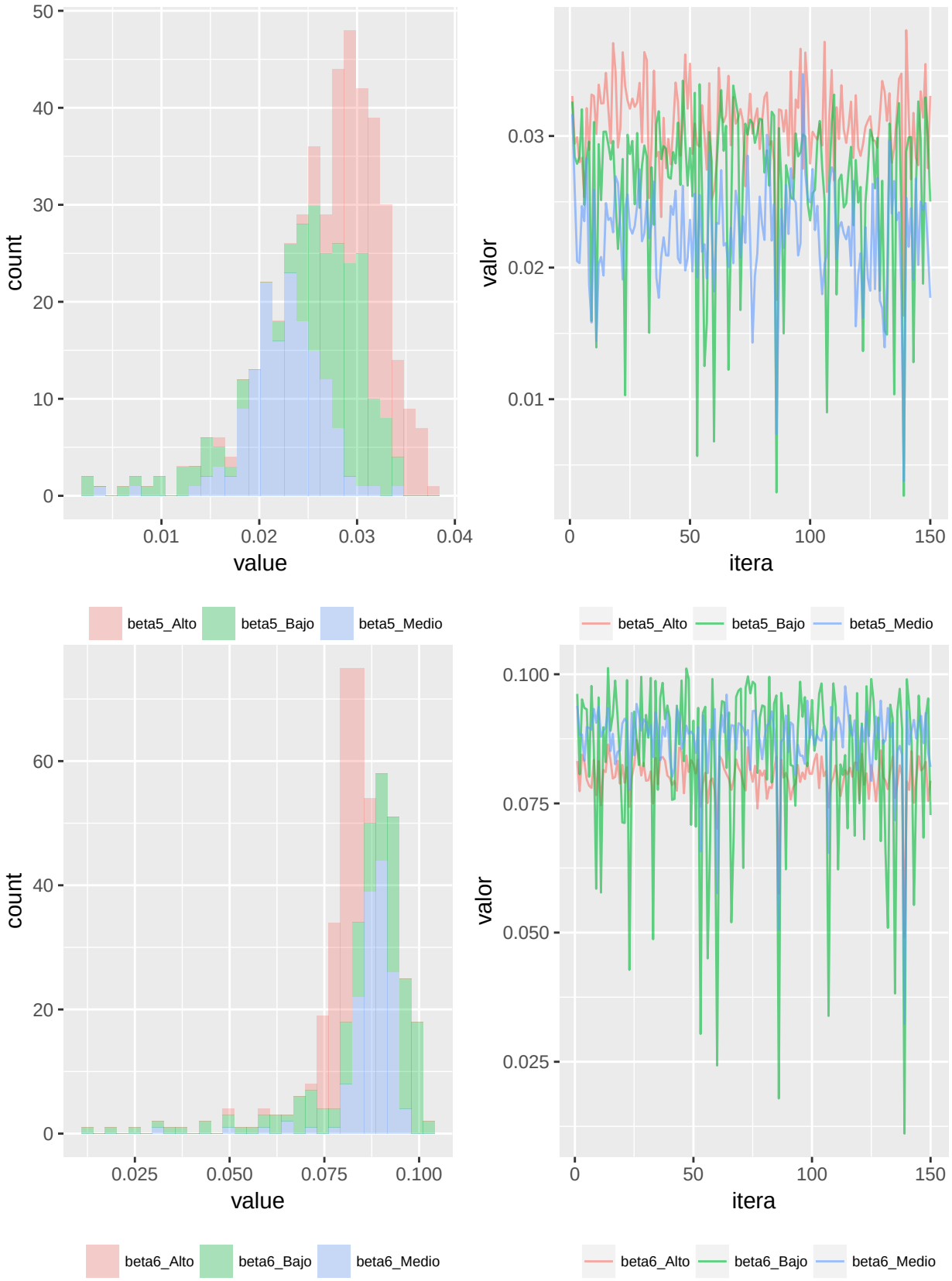
con $j = 1, 2, 3$ los segmentos Alto=1, Medio=2 y Bajo=3.

En este caso también las cadenas covergen de forma adecuada aún con pocas iteraciones como se aprecia en las siguientes gráficas para α_j y β_{1j}




count
value
valor
itera
beta1_Alto
beta1_Bajo
beta1_Medio
beta2_Alto
beta2_Bajo
beta2_Medio


count
value
1e−04
2e−04
3e−04
4e−04
0
20
40
60
80
valor
itera
0
50
100
150
beta3_Alto
beta3_Bajo
beta3_Medio
beta4_Alto
beta4_Bajo
beta4_Medio
0.01
0.02
0.03
0.04
0.05



Notamos que todos los parámetros son significativos pues en sus intervalos al 95% no contienen al cero.

Por otro lado, notamos que la diferencia de grupos se aprecia principalmente en las variables cardholder

y balance. En menor medida se aprecia dicha diferenciación en las variables numTrans, numIntTrans y creditLine.

De igual manera realizamos una tabla para la clasificación:

	0	1
0	0.966719	0.2462391
1	0.033281	0.7537609

Notamos resultados muy similares al de efectos independientes; es decir, que nuestro modelo tiene una especificidad alta; no obstante, la sensibilidad de 75% es baja, pues el efecto de tener un fraude es muy relevante.

Ahora analizaremos la clasificación por cada segmento. Primero se muestra el segmento alto:

	0	1
0	0.8068237	0.2210668
1	0.1931763	0.7789332

Ahora veamos el segmento medio:

	0	1
0	0.92663	0.3522907
1	0.07337	0.6477093

Finalmente, para el segmento bajo:

	0	1
0	0.9852097	0.2914286
1	0.0147903	0.7085714

El resultado anterior nuevamente resulta consistente, pues sabemos que en la categoría Alta al haber más fraudes, la predicción va a ser mejor. Por su parte, en las otras dos categorías como el porcentaje de fraudes es menor, no hay suficientes "éxitos" para generar una buena predicción.

Como constatamos el modelo de efectos independientes y el de intercambiabilidad que "comparte" la mayor cantidad de información entre segmentos arrojan resultados muy parecidos.

Hipótesis e interpretación de los modelos

Como el modelo de efectos independientes y el de intercambiabilidad son muy similares, en esta sección utilizamos solamente el segundo.

La primer hipótesis es que las personas con un número alto de transacciones (numTrans y numIntTrans) tienen mayor propensión a sufrir un fraude. De ser el caso se esperaría que los parámetros relacionados con estas dos variables sean distintos para cada segmento. El parámetro β_4 corresponde a numTrans y β_5 a numIntlTrans.

	mean	X2.5.	X97.5.
alpha[1]	-7.3550891	-7.8623733	-6.6322414
alpha[2]	-7.7687391	-8.3744390	-5.9019619
alpha[3]	-7.2456210	-8.3649526	-3.8835415
beta4[1]	0.0402741	0.0370044	0.0425096
beta4[2]	0.0441290	0.0336399	0.0467320
beta4[3]	0.0420678	0.0151020	0.0507761
beta5[1]	0.0304670	0.0245015	0.0361769
beta5[2]	0.0236995	0.0155695	0.0297637
beta5[3]	0.0256695	0.0081416	0.0338594

En el caso de las transacciones internacionales, el parámetro es distinto entre los segmentos, especialmente para el segmento alto. Los intervalos de confianza se empalman solamente en los valores altos pero hay diferencia entre estos. Estas características nos sugieren que en efecto el número de transacciones internacionales impacta en el número de fraudes diferenciando al segmento alto de los otros dos.

Por su parte, las transacciones domésticas β_4 tienen un efecto mayor en el segmento medio que en los otros dos. El efecto más débil es para el segmento alto. Considerando los intervalos de confianza podemos concluir que las diferencias de estos efectos entre un segmento y otro son menores que para las transacciones internacionales.

En la segunda hipótesis se plantea que las mujeres tienen una mayor predisposición al fraude que los hombres. El parámetro asociado es β_1 y la variable dummy es 0 para mujeres y 1 para hombres.

	mean	X2.5.	X97.5.
alpha[1]	-7.3550891	-7.8623733	-6.6322414
alpha[2]	-7.7687391	-8.3744390	-5.9019619
alpha[3]	-7.2456210	-8.3649526	-3.8835415
beta1[1]	-0.4992759	-0.6043764	-0.4034422
beta1[2]	-0.6602654	-0.7900899	-0.4753387
beta1[3]	-0.4551267	-0.6534186	-0.1218057

En los tres segmentos el parámetro es negativo indicando que si el cliente es hombre el valor estimado de fraude disminuye respecto a una mujer. En el segmento bajo la disminución es menor y para el segmento medio es la más alta. Esto confirma la hipótesis pues las mujeres obtienen valores más altos en el modelo. Los intervalos de confianza son similares y el efecto de esta variable en los segmentos es similar.

La tercer hipótesis es que los titulares de las tarjetas son más propensas a fraude que las adicionales. El párametro asociado es β_2 y como vemos en el primer segmento es positivo y en los demás negativo.

	mean	X2.5.	X97.5.
alpha[1]	-7.3550891	-7.8623733	-6.6322414
alpha[2]	-7.7687391	-8.3744390	-5.9019619
alpha[3]	-7.2456210	-8.3649526	-3.8835415
beta2[1]	0.0656823	-0.2563330	0.4272419
beta2[2]	-0.5492156	-0.9289340	-0.2101159
beta2[3]	-0.5119945	-0.8749295	-0.1518426

A diferencia de la variable género, en este caso el efecto de la variable en la estimación del fraude tiene distinto signo en los segmentos. Esto nos indica que si la tarjeta es del titular, el valor estimado de fraude aumenta en el primer segmento pero disminuye en los otros dos. Es importante considerar los intervalos

de confianza pues en el segmento alto el parámetro no es significativo. Además, este parámetro podría ser negativo y la conclusión se modificaría. Para el segmento medio y el bajo los intervalos de confianza son muy similares indicando que el efecto de la titularidad de la tarjeta es similar en ambos segmentos.

Predicción

Finalmente, para realizar las predicciones segmentaremos la muestra en 70% para entrenamiento y 30% para prueba y utilizaremos el modelo con efectos intercambiables.

	0	1
0	0.9374008	0.9363992
1	0.0625992	0.0636008

Notamos que las predicciones realizadas con el modelo jerárquico lineal generalizado con efectos intercambiabes es bastante bueno, pues tiene un precisión del 96% y una sensibilidad del 78%.

Ahora analizaremos las predicciones en cada uno de los segmentos:

Alto:

	0	1
0	0.6056653	0.6
1	0.3943347	0.4

En el segmento alto, el modelo arroja muy buenas predicciones, logrando una sensibilidad del 80%

Medio:

	0	1
0	0.8983193	0.8923077
1	0.1016807	0.1076923

El el segmento medio es donde se tiene la peor predicción, lo que se podría atribuir a que en este segmento hay un menor número de observaciones.

Bajo:

	0	1
0	0.9827063	0.9789343
1	0.0172937	0.0210657

Finalmente, como el segmento bajo tuvo una predicción aceptable con una precisión del 98% y una sensibilidad del 74%.

Conclusiones

La detección de fraude es díficil por el bajo número de ocurrencias. La proporción de fraude en los datos analizados es apróximadamente 6.6%. A partir de las variables balance y línea de crédito se generaron tres segmentos de nivel de fraude. El segmento alto concentra casi el 58% de los fraudes y en este es más fácil detectarlos. El segmento medio y el bajo tienen cerca de 20% de los fraudes cada uno pero el segmento medio tiene menos clientes por lo que su detección es más sencilla.

Tomando en cuenta los segmentos y el resto de las variables (excluyendo estado) se construyeron tres modelos. El primero es un modelo beta-bernoulli pero los resultados obtenidos no son buenos (sensibilidad de 33%). El segundo modelo es generalizado con efectos independientes y el tercero es un modelo generalizado con efectos intercambiables. Ambos convergen y arrojan resultados similares (sensibilidad cercana a 75%) y se seleccionó el modelo de efectos intercambiables.

De acuerdo con el modelo generalizado de efectos intercambiables las variables número de transacciones internacionales y titularidad de la tarjeta tienen parámetros distintos para los segmentos, especialmente el alto. Por su parte las variables número de transacciones domésticas y género no diferencían entre los segmentos. Con esta información podemos concluir que aquellos clientes del segmento alto titulares de la tarjeta y con alto número de transacciones internacionales son más propensos a fraude.

Las predicciones obtenidas con el modelo son muy buenas, logrando en general una precisión del 96% y una sensibilidad del 98%. La mejor predicción se obtuvo en el segmento "Alto", alcanzando una sensibilidad del 80% y la peor predicción se obtuvo en el segmento "Medio" con una precisión del 68%, lo cual se podría atribuir a que este segmento es el que tiene el menor número de observaciones.

Referencias

- Notas de clase del profesor Juan Carlos Martínez Ovando, en particular lo referente a modelos jerárquicos y modelos de regresión.
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. Bayesian Data Analysis, 2002, 2a edición. Chapman & Hall: Boca Raton.
- Gelman, A., Hill, J. Data Analysis Using Regression and Multilevel / Hierarchical Models, 2008, 6a edición, Cambridge University Press.