1. **Introducción**

**1 párrafo**

1. **Datos**

|  |  |
| --- | --- |
| **Variable categórica** | **Característica** |
| Oficio |  |
| MaxNivEdu |  |
| TipoDeViv |  |

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

P5010 (cuartos usados)

NPer

Fex\_c

Fex\_depto

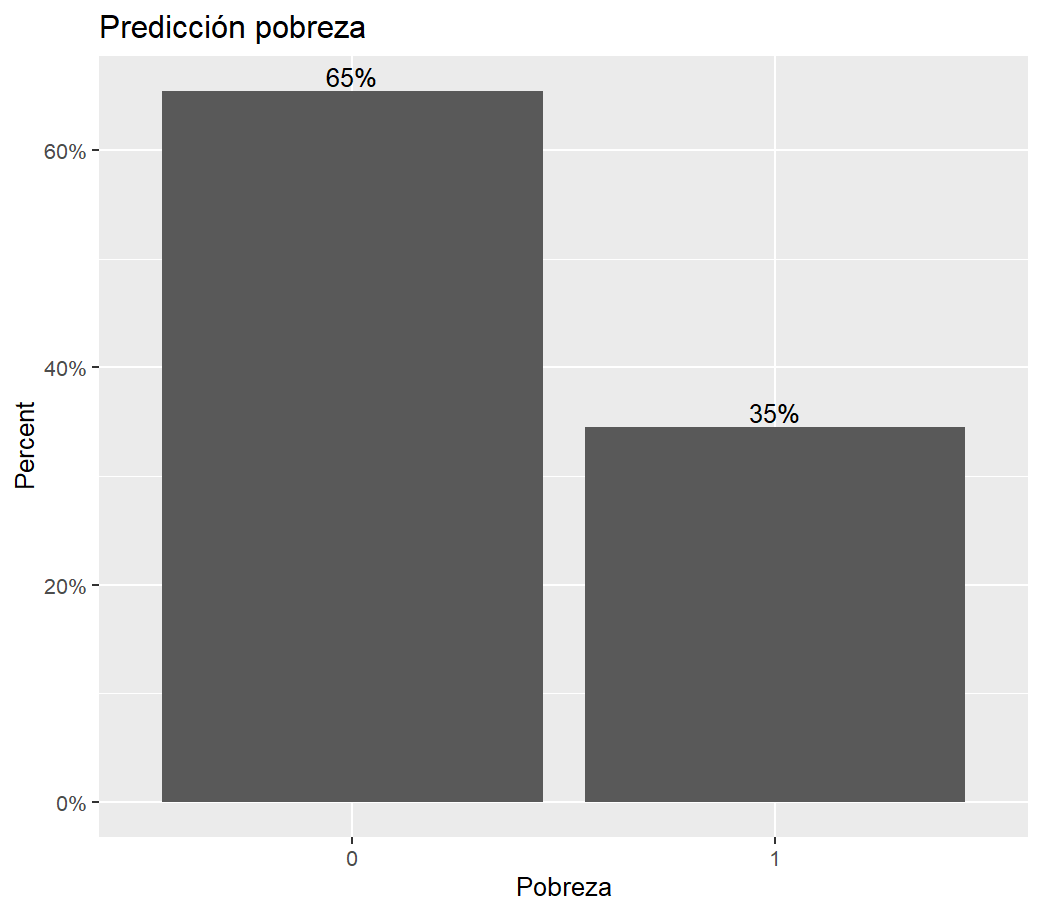
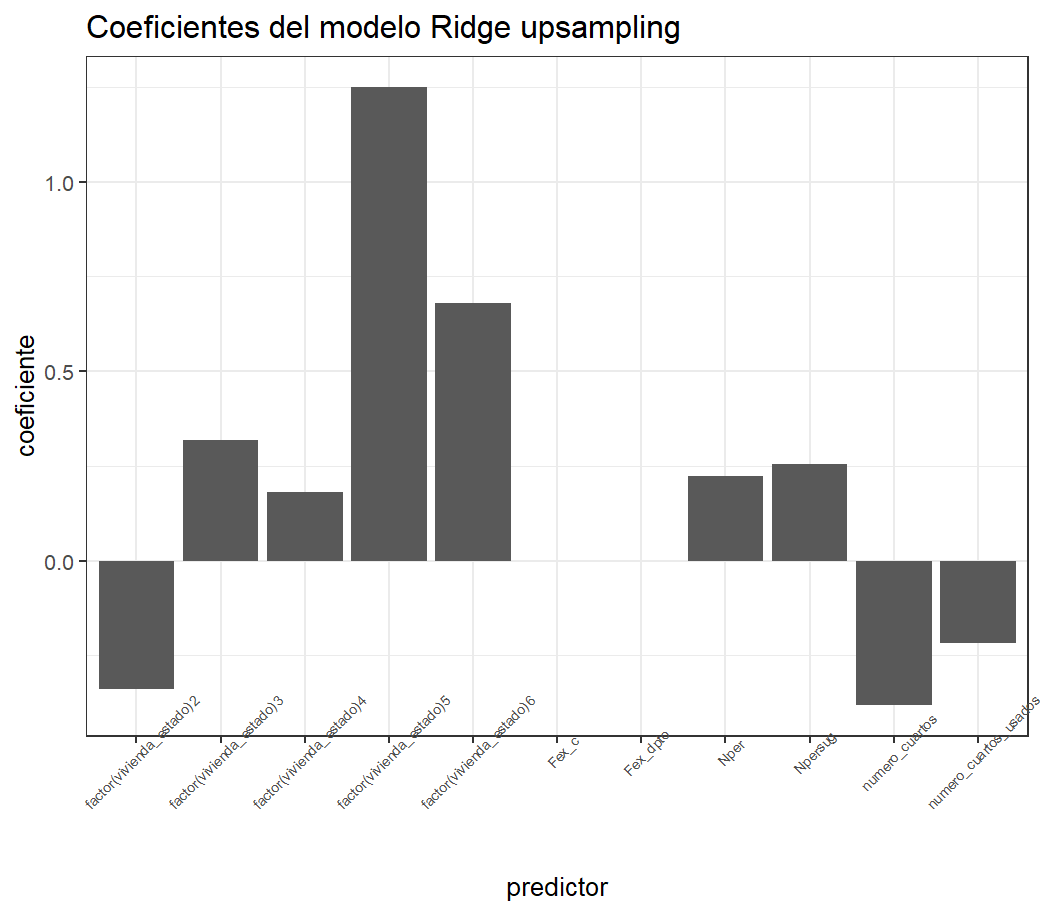
**3.Modelos y resultados**

**3.1 Modelo de clasificación**

Forma funcional del modelo: **Esta misma forma funcional se utilizó para los 7 modelos realizados en el procedimiento de clasificación.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelos** | | **Lambda mínimo** | **AUC** | **Tasa Falsos positivos** | **Sensitivity** |
| Modelo 1 | Estimación con Lasso – Alpha=1 | 0,0006911766 | 0,5714028 | 16,9% | 16,8% |
| Modelo 2 | Estimación con Ridge – Alpha=0 | 0,00957201 | 0,550705 | 17,9% | 11,73% |
| Modelo 3 | Estimación Logit – Lambda=0 | 0 | 0,5747561 | 16,7% | 17,81% |
| Modelo 4 | Estimación con Lasso-Upsampling | 0,0006887055 | 0,6892392 | 7,22% | 64,54% |
| Modelo 5 | Estimación con Ridge-Upsampling | 0,0138377 | 0,6891904 | 7,21% | 64,62% |
| Modelo 6 | Estimación con Lasso-Downsampling | 0,0009165909 | 0,689228 | 7,27% | 64,31% |
| Modelo 7 | Estimación con Ridge-Downsampling | 0,01393137 | 0,6888365 | 7,25% | 64,40% |

El mejor modelo de predicción de la pobreza a partir del procedimiento de clasificación fue el Modelo 5 – Estimación Ridge con Upsampling de la muestra. Este es el mejor modelo ya que, en primer lugar, es el modelo con menor cantidad de falsos positivos (7,21%) lo cual se puede interpretar como un buen modelo predictivo de la pobreza al ser este porcentaje bajo. Así mismo, es el modelo con mayor Sensibilidad (64,62%), es decir, la proporción de hogares pobres que efectivamente fueron clasificados como pobres. Por último, es el modelo que también tiene mayor AUC (0,6891904) lo que se puede interpretar como el modelo con mayor poder predictivo.



La Gráfica de la derecha no muestra un histograma de las predicciones del modelo, analizándose que predijo que el 35% de los hogares son Pobres. La Gráfica de la izquierda nos muestra la importancia de cada variable independiente en la predicción de la pobreza, determinándose que la variable explicativa que más impacto tiene en la predicción de la pobreza de los hogares es el estado de la vivienda, específicamente, los hogares que tienen vivienda a partir de un usufructo o una posesión sin título tienen una mayor probabilidad de ser pobres frente a los hogares que si tienen un hogar propio. Así mismo, la gráfica nos muestra que el número de las personas en el hogar si es determinante para predecir la pobreza en los hogares, específicamente, entre mayor sea el número de personas en la unidad de gasto del hogar la probabilidad de que sea pobre aumenta significativamente.

**3.3 Modelo de regresión**

En estos 3 primeros modelos se utiliza a base de datos personas, con el fin de obtener la predicción de cada uno de los individuos para finalmente sumarlos por hogar compararlo en la base de datos trainhogar.

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | | **MSE** | **Mejor lambda** |
| Modelo 1 |  | 1.058E+12 | No aplica |
| Modelo 2 | Mismo modelo 1 estimado con Ridge | 1.060E+12 | 49197 |
| Modelo 3 | Mismo modelo 1 estimado con Lasso | 1.058E+12 | 239 |
| Modelo 4 |  | 5.888E+12 | No aplica |
| Modelo 5 | Mismo modelo 4 estimado con Ridge | 5.891E+12 | 91457 |
| Modelo 6 | Mismo modelo 5 estimado con Lasso | 5.890E+12 | 3443 |
| Modelo 7 |  | 1.058E+12 | 239 |

**4.Conclusiones**

Este ejercicio predictivo de la pobreza de los hogares colombianos nos muestra que la forma para predecir una variable puede ser a partir de diversos procedimientos, siendo de suma importancia desarrollar diversos modelos para compararlos entre si y encontrar el que mejor prediga.

La mayor conclusión de la predicción de la pobreza a partir del problema de clasificación es que el remuestreo es una forma eficiente para que nuestros modelos hayan aumentado su poder predictivo y disminuido sustancialmente los falsos positivos a comparación de los modelos sin remuestreo. En este caso en específico, nuestra muestra presentaba un desequilibrio significativo en la variable “Pobre” por lo que esto podía estar impactando negativamente la efectividad de los modelos.

Nuestro análisis nos llevó a determinar que el mejor modelo de predicción para la pobreza de los hogares a partir de nuestra forma funcional es un ridge con upsampling, estando los demás modelos con remuestreo muy cerca del poder predictivo de este modelo. Algo a destacar de este modelo es que los falsos positivos encontrados fueron de 7,21%, una cifra muy baja frente a la predicción de una variable de alta complejidad como lo es la pobreza de los hogares.