Para modelos de clasificación se optó por un modelo Logit Ridge con *upsampling.* Este fue el modelo que arrojó los mejores resultados sobre sensibilidad. Teniendo en cuenta que el objetivo principal es evitar los falsos negativos y que queden personas pobres sin el subsidio. Los primeros tres modelos evaluados dan una sensibilidad de 1 que es “perfecta” y no presentan falsos negativos, pero estos son modelos *naïve* queclasifican a todos como “pobres” por lo que se evita el problema del falso negativos. Por lo tanto, no presentan un verdadero valor predictivo. Una explicación del buen desempeño de este modelo es el *up sampling* que se realiza. La base de train tiene una proporción de 5.8 % de pobres en la muestra. Esta metodología simula nuevas observaciones de la minoría (pobres en este caso) en la muestra para poder tener más información para realizar la predicción. Por último, este modelo tiene en cuenta todas las variables creadas y tuvo el mejor resultado a pesar de la penalización por número de variables que presenta un modelo Ridge.

Los hiperparámetros que maximizan la sensibilidad para este modelo son (característico de un modelo Ridge) y .

A continuación se presenta una tabla comparativa entre los modelos estudiados:

Modelos de regresión.

El modelo con mejor MSE es el siguiente:

Para este modelo se utilizó un *down sample,* es decir, se eliminan de forma aleatorios valores de “no pobres” en la muestra para tener una base más balanceada. Este modelo utilizó todas las variables que se tenían para el modelo, incluyendo interacciones y variables categóricas por ciudad. Se debe tener en cuenta que, aunque el modelo presenta el mejor resultado de la variable deseada, éste sigue siendo un resultado bastante alto. Los modelos sin tener ajustes al sampleo mostraron resultados peores de predicción, lo que indica la importancia de rebalancear la muestra para entrenar un modelo.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelos** | **RMSE** | **MSE** | **Rsquared** | **MAE** |
| DownSample Completa | 1.803.241 | 3,2517E+12 | 0,2212784 | 9394759 |
| Upsample Completa | 1.845.479 | 3,4058E+12 | 0,207994 | 918369 |
| Regresión sin resampleo | 2.309.201 | 5,3324E+12 | 0,1426602 | 1253953 |
| Upsample y preprocesamiento | 1.849.596 | 3,421E+12 | 0,2044551 | 920673 |
| Regresión sin resampleo y con pre procesamiento | 2.302.387 | 5,301E+12 | 0,1476736 | 1249887 |