Problem Set 2: Predicting Poverty

Big Data y Machine Learning para Economía Aplicada

Gustavo Adolfo Castillo Álvarez (201812166),
Alexander Almeida Ramírez (202225165),
Jorge Luis Congacha Yunda (201920042) y
Jaime Orlando Buitrago González (200612390)

2024-04-14

1 Introducción

2 Datos (al final)

2.1 Procesamiento de Datos

2.1.1 Consolidación de base "clean"

2.1.2 Construcción de variables (feature engineering)

2.1.3 Ejercicio con variables DANE

2.2 Análisis Descriptivo

3 Modelos y resultados

3.1 Modelos de Clasificación

3.1.1 Enfoque

Para clasificar los hogares se buscó directamente predecir la clase $k \in \{Pobre, NoPobre\}$, a la que pertenecen los hogares, en este sentido se trató de un problema de clasificación binaria pues solo hay dos clases posibles. Dado que un hogar es clasficado como pobre directamente si su ingreso es menor a una línea de pobreza, se buscó estimar la categoría de un hogar a partir de características socioeconómicas del hogar: proporción de mujeres, edad promedio de los integrantes del hogar, número de menores, etc. Así, el objetivo era aproximar la probabilidad condicional de que un hogar con cierto vector de características $\boldsymbol{X}$ fuera pobre o no, i.e. $\mathbb{P}[y_i = k|X_i = x] = p(x)$. La expresión analítica de $p(x)$ depende del modelo estimado. En las dos regresiones logísticas, con y sin regularización lasso, se preprocesaron estandarizandolas. En todos los modelos realizados se usaron 15 variables independientes: edad media del hogar, edad media al cuadrado, proporción de mujeres, proporción de estudiantes, proporción de personas buscando trabajo, proporción de personas dedicadas a oficios del hogar, proporción de personas con primaria, prop. de personas con secundaria, prop. de personas con educación superior, número de cuartos, prop. de personas en régimen contributivo, número de adultos mayores, dicotómica de hogar rural, y dicotómica de si la vivienta es arrendada.

Utilizamos 6 diferentes algoritmos de clasificación: logit, QDA, LDA, *random forest*, *bagging* y *GBM*. Dado que cada uno de estos modelos "aprende" la frontera de decisión utilizando diferentes enfoques sus ecuaciones respectivas varían significativamente. El único modelo discriminativo que implementamos, el logit, directamente estima la probabilidad condicional mediante una función logística (1).

$$\mathbb{P}[y_i|X_i] = \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}\boldsymbol{\beta})} \tag{1}$$

$$\mathbb{P}[y_i = k|X_i = x] = \frac{\pi_k f_k(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}, k \in \{0, 1\} \tag{2}$$

Este modelo obtuvo un valor de $F1$ de

Los dos modelos generativos como el QDA y LDA en cambio descomponen $p(x)$ como una distribución posterior y mediante Bayes y haciendo supuestos sobre la densidad condicional $p(x, y)$ la estiman de diferentes formas, en ambos casos suponiendo que surgen de distribuciones Gausianas mutivariadas. En el caso de LDA supone una varianza común entre las clases, mientras que QDA supone una varianza y media específica a cada clase.

Los últimos modelos basados en árboles posibilita unas relaciones no lineales entre las características y la variable dicotómica de Pobre. Tanto el *Random Forest* como el *bagging* agregan varios árboles de forma independiente usando bootstrap en cada nuevo árbol, utilizando el primero un subconjunto alreatorio de los predictores en cada iteración.

3.1.2 Ecuaciones

adfad

3.2 Modelos de Predicción de Ingreso

3.3 Modelos Finales (QDA)

4 Conclusiones

5 Referencias bibliográficas