Modelos de clasificación para determinar declaración de renta sobre muestras con clases desbalanceadas

El caso del impuesto sobre la renta en Colombia

Julián Enrique Chitiva Bocanegra
201317222

Paula Rodríguez Díaz
201327494

***Resumen*—En los países de América Latina existe un déficit fiscal debido al bajo recaudo de impuestos. Este artículo examina el caso de impuestos sobre la renta (impuesto sobre el ingreso) en Colombia prediciendo, con base en características socio demográficas, si un ciudadano debería o no declarar renta.**

Palabras clave*—*impuestos, renta, evasión, predicción, imbalanceo de clases, aprendizaje supervisado.

I. Introducción

Durante la última década, el gobierno colombiano ha realizado cerca de 10 reformas tributarias las cuales han buscado aumentar el recaudo tributario para reducir brechas sociales y económicas. Se estima que la evasión de impuestos corresponde a más de 40 billones de pesos anuales, la cual se debe principalmente a que los contribuyentes no tienen incentivos a tributar por desconfianza en las instituciones que administran estos recursos.

El motivo de este trabajo es presentar una metodología para mejorar el recaudo de impuestos, con base en variables socio demográficas que sean observables para el gobierno, basándonos en la idea que personas socialmente parecidas deben tener niveles de tributación similares.

Para nuestras estimaciones usamos los datos de la Encuesta de Calidad de Vida (ECV) realizada por el Departamento Administrativo Nacional de Estadística (DANE).

II. Descripción de los datos

Se realizaron las estimaciones con la ECV para los años 2015 y 2016, debido a que esta encuesta es un referente internacional en cuanto a su metodología y ejercicio estadístico. Por lo tanto, es una muestra representativa de la población colombiana. La metodología es similar a las implementadas por el Banco Mundial para medir condiciones de vida. Además de esto, las encuestas dirigidas a los hogares son una de las fuentes más importantes de variables sociales, económicas y demográficas. Con estos indicadores el DANE calcula indicadores económicos que usamos para estimar si un individuo debe declarar renta o no.

Con la ECV el DANE encuesta en promedio 20.000 hogares que cuentan con aproximadamente 50.000 adultos. De estos, tan solo el $3\,\%$ son los que por sus ingresos anuales deben declarar renta. El promedio de edad de las personas en la muestra es de $44\,\%$ años y estas personas en promedio viven en estrato 2. Estas características muestran que la base de datos es representativa y en especial, la variable de interés (renta) es acorde con la realidad colombiana. Según Londoño y Alvaredo (2013, p. 35) el promedio de personas que debe declarar renta

en Colombia está entre el $1,5\,\%$ y el $5\,\%$.

Los modelos que estudiaremos serán entrenados con la base de datos de la ECV de los años 2015 y 2016. Contamos con 246 variables de las cuales 214 son categóricas y 32 son numéricas continuas. Se tomaron aquellas variables de la ECV que fueran equivalentes entre ambos años tanto para la pregunta realizada en la encuesta como los niveles de respuesta posibles. Habiendo eliminado las observaciones para menores de edad terminamos con una base de entrenamiento con 103.534 observaciones. Para la validación de los modelos utilizamos la ECV del año 2014 para estudiar la generalización de los modelos hacia otros años.

III. Modelo

En las bases de datos las variables que contienen información sensible como el ingreso de las personas, normalmente presentan sesgo al ser auto reportadas. Por tal razón, se realizó la siguiente transformación de las variables. La ECV tiene una estimación del ingreso por hogar y una variable de salario mensual (auto reportado) y por los sesgos antes mencionados, se definió una nueva variable de ingreso como el máximo entre el ingreso por adulto en el hogar (dividir el ingreso por hogar entre el número de adultos) y el salario. Se decidió hacer esto debido a que, si bien existen menores de edad que pueden contribuir al ingreso del hogar, la baja probabilidad de declarar renta que tiene un hogar con estas condiciones.

Para definir si un hogar debe declarar renta se decidió usar solo la condición sobre el ingreso, debido a que las otras condiciones como el patrimonio, el consumo con tarjeta de crédito o el acumulado de consignaciones y depósitos son más observables para el gobierno. En la tabla III se muestran el ingreso anual que obliga a los ciudadanos a declarar renta.

Año gravable	Ingresos mayores a
2014	38.479.000
2015	39.591.000
2016	41.654.000

Con esta nueva variable se consideró el siguiente modelo:

$$renta_i = f(\boldsymbol{X}_i) + \varepsilon_i \tag{1}$$

donde $renta_i$ toma valores en $\{0, 1\}$, y $\boldsymbol{X}_i$ son variables socio demográficas del individuo i.

IV. Algoritmo

Los algoritmos seleccionados para predecir si una persona debe declarar renta o no fueron CART, boosting de árboles y *Naive Bayes* (Bayes ingenuo). En la base de datos original los declarantes representan muy poca parte de la población, por lo cual hay un gran desbalanceo entre las clases consideradas. Los modelos de clasificación sobre muestras desbalanceadas suelen tener bajos desempeños pues tienden a favorecer la clase mayoritaria. Por consiguiente, se decidió hacer balancear las clases haciendo un remuestreo y submuestreo siguiendo las técnicas de sobre muestreo sintético sobre minorías (SMOTE) propuesto por Chawla *et al.* (2011). Posteriormente se usaron estos nuevos datos para entrenar los modelos antes mencionados.

IV-A. SMOTE

Los modelos SMOTE (en inglés "Synthetic Minority Over-sampling Tecnhique"), buscan hacer un sobre muestreo sobre clases minoritarias (los declarantes de renta) creando ejemplos sintéticos antes que hacer muestreo con reemplazo. Este sobre muestreo se realiza escogiendo puntos aleatoriamente entre los intervalos de cada variable de la clase

minoritaria y posteriormente se adicionan los k-vecinos más cercanos dentro de la minoría. Este procedimiento fuerza a la región de decisión a ser más general dentro de la clase minoritaria. De esta manera, los modelos de regresión y clasificación aprenden más de la clase minoritaria en lugar de ésta ser opacada por los datos en la clase mayoritaria a su alrededor, logrando que los modelos que usan árboles generalicen mejor. Adicionalmente el algoritmo SMOTE realiza un submuestreo de la clase mayoritaria eliminando aleatoriamente hasta que se alcance la proporción deseada de los datos. Además, el paquete 'DMwR' de R permite llevar a cabo este algoritmo teniendo variables categóricas, lo cuál fue bastante conveniente dada la naturaleza de los datos trabajados.

IV-B. CART

Los modelos CART consisten en árboles de clasificación y regresión; sin embargo, debido a la naturaleza de la variable de interés se usarán los de clasificación binarios. Estos métodos se basan en árboles que parten el espacio de las variables predictores en rectángulos n-dimensionales y asignan como respuesta el promedio de ésta dentro de cada partición. Para definir las regiones se realiza un *greedy algorithm* que permite seleccionar el problema de manera computacionalmente factible. Dado que la mayoría de las variables en cuestión son categóricas, el modelo CART tendería a funcionar bastante bien por las particiones que realiza.

IV-C. Gradient Boosting de árboles

El gradient boosting de árboles es un algoritmo que reduce el sesgo y la varianza en el contexto de aprendizaje supervisado. Este consiste en combinar resultados de clasificadores débiles para obtener un clasificador robusto. En este método, se crean árboles secuencialmente usando la información y aprendizaje de los anteriores. Este procedimiento no realiza *bootstrap* sobre la muestra, en lugar de esto ajusta cada árbol en una versión modificada de los datos.

IV-D. Naive Bayes

Este método consiste de clasificadores sencillos basados en la aplicación de la regla de Bayes con fuertes suposiciones de independencia entre las características. Este algoritmo a pesar de su simplicidad puede llegar a competir con los anteriormente mencionados.

V. Aplicaciones a la base de la encuesta de calidad de vida

Para predecir si un ciudadano colombiano debe o no declarar renta según sus características socio demográficas entrenamos los modelos CART, Gradient Boosting y Naive Bayes tanto para la muestra desbalanceada como la muestra balanceada usando el algoritmo SMOTE. Es interesante observar los diferentes modelos que se obtienen cuando las muestras están balanceadas y desbalanceadas. Por ejemplo, para el modelo CART obtuvimos las siguientes particiones dada la muestra desbalanceada y balanceada por SMOTE.

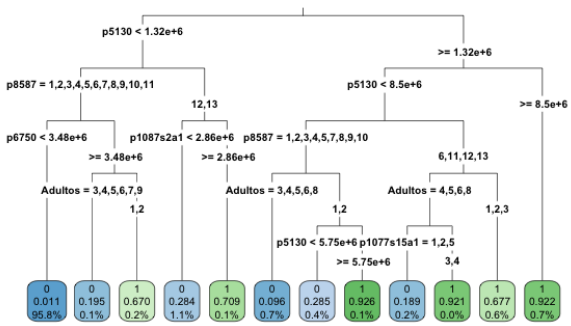

Figura 1. Árbol CART entrenado con muestra desbalanceada

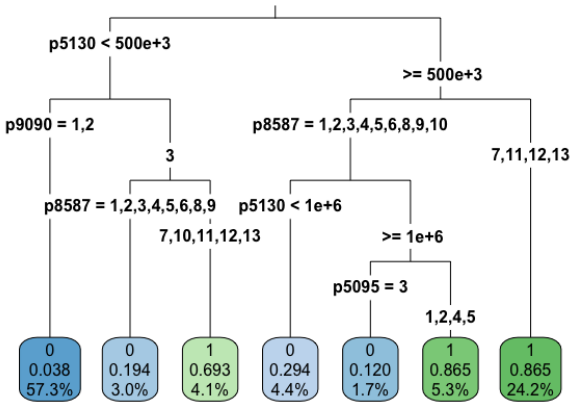

Figura 2. Árbol CART entrenado con muestra balanceada por SMOTE

Es evidente que los modelos obtenidos son completamente distintos. Por lo tanto, es importante determinar si el desempeño de los modelos efectivamente mejora o empeora al ser entrenados con clases balanceadas por SMOTE. Para la evaluación de los modelos observamos sus curvas ROC correspondientes y su valor de desempeño AUC (área bajo la curva ROC).

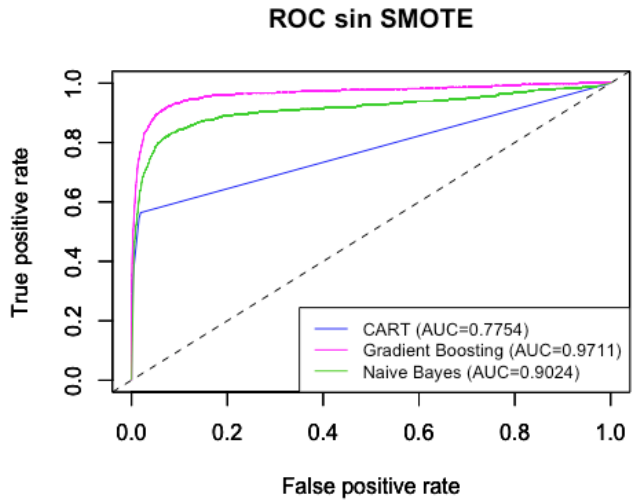

Figura 3. Curva ROC para los modelos entrenados sin balancear clases

En la Figura 3 se observa que el desempeño de CART entrenado con la muestra desbalanceada no es muy bueno; tiene un $AUC = 0.7754$. Sin embargo, el desempeño de Gradient Boosting (GB) entrenado con la muestra desbalanceada es bastante bueno; tiene un $AUC = 0.9711$. Aunque el modelo GB funciona por medio de árboles como CART, el hecho de hacer un ensamblaje de estos (2000 árboles) hace que el modelo tenga un muy buen desempeño. Esto seguramente se debe a que la mayor parte de los predictores son categóricos y además la base de datos considerada tiene predictores son varios datos faltantes.

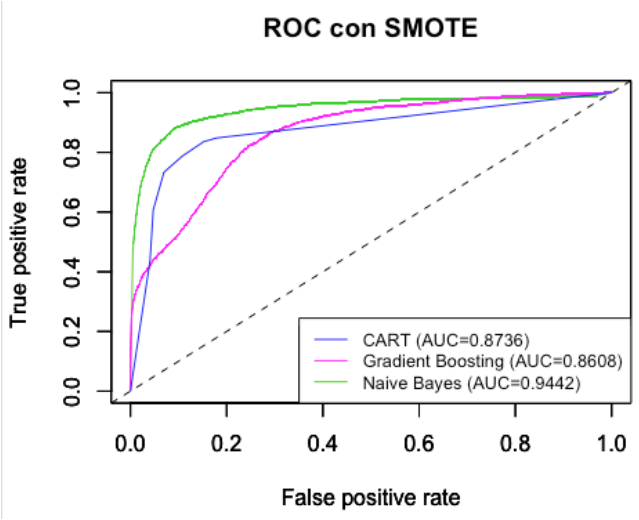

Figura 4. Curva ROC para los modelos entrenados con clases balanceadas por algoritmo SMOTE

En la Figura 4 se observa que el desempeño de CART aumentó considerablemente al ser entrenado con la muestra balanceada por SMOTE. El AUC paso de ser 0.7754 a 0.8736. Sin embargo, el desempeño de GB empeoró, su AUC paso de ser $0.0.9711$ a 0.8608. El desempeño de Naive Bayes mejoró al entrenarlo con la muestra balanceada, su AUC paso de ser $0.0.9024$ a 0.9442.

VI. Conclusiones

Pudimos observar que clasificadores sencillos como Naive Bayes generalizan muy bien teniendo así un buen desempeño al clasificar los ciudadanos en declarantes o no declarantes para las encuestas del 2014 con la mayor parte de los predictores siendo

categóricos. Además, su desempeño mejora considerablemente cuando éste se entrena con una muestra balanceada por medio del algoritmo SMOTE. Sin embargo, para modelos más complejos y robustos como Gradient Boosting de árboles de clasificación, encontramos que su desempeño fue mejor cuando se entrenó con la clase desbalanceada; incluso, éste fue el mejor modelo de los seis modelos estudiados. Encontramos entonces que GB maneja muy bien el problema de clases desbalanceadas y además generaliza muy bien la clasificación deseada de año a año. Acercamiento como éste para predecir si un ciudadano debe o no declarar renta por medio de variables socio demográficas y, teniendo en cuenta el bajo porcentaje de personas que en realidad debe hacerlo, puede ser de gran importancia para el gobierno para así poder evitar la evasión del impuesto a la renta.

Referencias

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

García, A. (2015). "los colombianos evaden $40 billones en impuestos al año": Director de la dian.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2103). *An Introduction to Statistical Learning - with Applications in R*. Springer.

Londoño, J. y Alvaredo, F. (2013). High incomes and personal taxation in a developing economy: Colombia 1993-2010". *Commitment to Equity Working Paper*, (12).