Regresión Logística para predicción de Rezago Social en México 2015

Pablo Rodrigo Ávila Solís (avila.pablo@colpos.mx)

Posgrado en Socioeconomía, Estadística e Informática – Cómputo Aplicado, Colegio de Postgraduados

Introducción

El rezago social es una condición que afecta a millones de personas que, además de estar correlacionado positivamente con la pobreza, debe ser medido con regularidad para poder ser combatido. Con la finalidad de tomar decisiones estratégicas adecuadas y garantizar su monitoreo continuo es importante contar con herramientas que permitan estimar el grado de rezago social en México. En este sentido, se propone una método para la clasificación del rezago social, a nivel municipal, con base en el conteo de las unidades económicas presentes en el área de interés. La clasificación del rezago social se abordó como un problema de aprendizaje automático supervisado mediante un modelo de regresión logística multinomial.

Objetivos

- Elaborar una método para clasificar el grado de rezago social municipal, con base en datos del Directorio Estadístico Nacional de Unidades económicas de 2015.
- Obtener un modelo base para comparar la clasificación de cinco categorías oficiales del grado de rezago social.

Materiales y Métodos

Las clases objetivo, fueron los cinco grados de rezago social reportados por Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), en las que se distribuyeron 2446 municipios en 2015. Por la naturaleza del fenómeno de rezago social, estas clases están desbalanceadas de la siguiente manera:

Clase	Frecuencia	Porcentaje
1 - Muy bajo	760	31.07%
2 - Bajo	341	13.94%
3 - Medio	603	24.65%
4 - Alto	567	23.18%
5 - Muy alto	175	7.16%
Total	*2446*	*100.0%*

Los datos de entrada fueron conteos de unidades económicas de acuerdo a las diferentes categorías del Sistema de Clasificación de América del Norte (SCIAN). Los niveles de las categorías del SCIAN son: sector, subsector, rama, clase y subclase. De esta manera, los distintos conteos se agruparon a nivel municipal y se dividieron entre la población del municipio, para obtener tasas como parte de las variables de entrada. Además, se agregaron como características de entrada las coordenadas geográficas de los centroides de los municipios.

Materiales y Métodos …

Se ajustó el modelo regresión logística multinomial que, para determinar la probabilidad de la variable dependiente, hace uso de la función $\sigma(z) = \frac{1}{1-e^{-z}}$, donde $z = w_0 + w_1x_1 + \ldots + w_px_p = \sum_{j=0}^{p} w_jx_j = \mathbf{w}^{\mathrm{T}}\mathbf{x}$, $x_0 = 1$, p es el número de variables de entrada, $\mathbf{w}$ es un vector de parámetros y $\mathbf{x}$ el vector de variables de entrada del modelo. La gráfica de la función anterior es una curva en forma de S, con valores de su domino en $\mathbb{R}$ y $(0,1)$ en su contradominio.

El modelo se regularizó para penalizar el número de variables de entrada y contraer los pesos que menos contribuyen hacia cero. De esta forma, la función de costo utilizada fue:

$$J(\mathbf{w}) = \sum_{i=1}^{n}\left[y^{(i)}log\left(\sigma(z^{(i)})\right) - \left(1-y^{(i)}\right)log\left(1-\sigma(z^{(i)})\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{p} w_j^2,$$

donde $y^{(i)}$ representa el valor verdadero de la i-ésima observación y $\sigma(z^{(i)})$ la predicción de $y^{(i)}$ para el mismo ejemplo. Esta modelación se realizó con la librería *Scikit-learn* de python en *Google Colab*.

Se probaron, por medio de búsqueda en malla, dos métodos de optimización para los parámetros $\mathbf{w}$ (*newton-cg* y *lbfgs*), distintos valores de tolerancia para el criterio de parada, y distintos valores del hiperparámetro de regularización λ. Los datos de entrada se estandarizaron durante el ajuste, con el fin de ayudar a la convergencia de las técnicas de optimización empleadas.

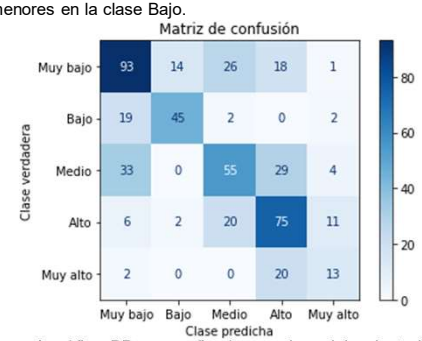
La estrategia de selección de parámetros e hiperparametros consistió en la técnica de validación cruzada, con 5 grupos, tomando el 80% (1956) de los datos para entrenamiento y 20% (490) para probar el rendimiento del modelo ajustado. Dado que los datos no están balanceados, todas las particiones fueron estratificadas.

Las métrica de desempeño que se optimizó durante el entrenamiento fue el valor *F1-macro*, en virtud de que esta medida es adecuada cuando la distribución de las etiquetas objetivo no es uniforme. Una vez elegido el modelo final, se calcularon, a partir del conjunto de prueba las métricas: *F1-macro*, precisión global (*accuracy*), precisión global balanceada (*balanced accuracy*), así como la matriz de confusión y curvas *Precision-Recall* (PR) para observar la predicción obtenida por cada clase de rezago.

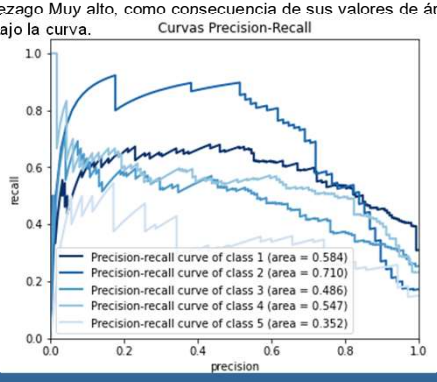
Cabe destacar que, las etiquetas se modelaron con base en un enfoque nominal, y no bajo la ordinalidad que guardan las clases de rezago social.

Resultados

Los hiperparámetros óptimos, bajo la métrica de desempeño *F1-macro*, durante el entrenamiento, fueron: tolerancia de 0.05, $\lambda = 0.001$, y el método *newton-cg* para la optimización de los pesos $\mathbf{w}$. Lo anterior, bajo el escenario de las variables de entrada a nivel subsector de acuerdo al SCIAN. Los valores de las métricas de desempeño en el conjunto de prueba fueron: F1-macro de 0.55, precisión global de 0.57 y precisión global balanceada de 0.55. La matriz de confusión mostró que, se presentaron mayores errores de clasificación en la clase Muy alto, y menores en la clase Bajo.



Con el gráfico PR se confirmó que el modelo ajustado es más preciso para clasificar el grado de rezago social Bajo y tiene un desempeño mediocre para predecir la clase de rezago Muy alto, como consecuencia de sus valores de área bajo la curva.



Resultados …

Considerando que, el clasificador nulo tiene una precisión general de 0.31 si se toma en cuenta la clase más frecuente, y de alrededor de 0.2 si se adivina al azar alguna de las cinco clases; se puede deducir que el modelo "aprende", con cierto sesgo, a clasificar las etiquetas de rezago social. Así, con base en los resultados, se sugiere tomar un enfoque ordinal para la modelación, o bien, aproximar el desbalanceo de las clases con métodos *SMOTE* de submuestreo o sobremuestreo. Asimismo, seleccionar variables, en lugar de usar todo el conjunto de datos, para evitar posible multicolinealidad. También, es factible enfocarse en alguna clase de rezago de interés para optimizar el aprendizaje supervisado sobre la misma.

Conclusiones

El desempeño del modelo de regresión logística multinomial regularizado, mostró que un clasificador lineal no es ideal para ajustar todas la clases de rezago social. Sin embargo, se obtuvo un modelo base con mejores resultados que la predicción aleatoria para la mayoría de las etiquetas, bajo el paradigma de aprendizaje automático supervisado considerando un enfoque nominal, con base en datos del Directorio Estadístico Nacional de Unidades Económicas a nivel subsector.

Referencias

- CONEVAL (Consejo Nacional para la Evaluación de la Política de Desarrollo Social). 2016. Índice de Rezago Social 2015. Presentación de Resultados. https://www.coneval.org.mx/Medicion/Documents/Indice_Rezago_Social_2015/Nota_Rezago_Social_2015_vf.pdf
- Géron, A. 2019. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2a ed. O'Reilly Media. Sebastopol, CA.
- INEGI (Instituto Nacional de Estadística y Geografía). 2013. Sistema de Clasificación. Industrial de América del Norte, México SCIAN 2013. https://www.inegi.org.mx/contenidos/productos//prod_serv/contenidos/espanol/bvinegi/productos/clasificadores/SCIAN/SCIAN_2013/702825051693.pdf
- INEGI (Instituto Nacional de Estadística y Geografía). 2014. Marco geoestadístico 2014 versión 6.2 (DENUE). https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825004386
- INEGI (Instituto Nacional de Estadística y Geografía). 2015. Directorio Nacional de Unidades Económicas DENUE. https://www.inegi.org.mx/app/mapa/denue/default.aspx
- James G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning: with applications in R. Springer.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12: 2825-2830.
- Raschka S. and V. Mirjalili. 2019. Python Machine Learning. 3a Ed. Packt Publishing Ltd. Birmingham, UK.