Balanceo en la GEIH

Wilson Andrés Pinzon

Introducción

En el siguiente documento se presenta un método que busca tratar el problema del desbalance de datos en machine learning a través de un estudio sobre la variable `ACTIVIDAD_OCUPADA_ULTIMA_SEMANA`, del conjunto de datos de la GEIH, la cual será la variable dependiente de un modelo de regresión logística que busca explicar qué impacto tienen algunas condiciones sociales y demográficas en la probabilidad de que un joven de 18 a 28 años que pertenece a la fuerza de trabajo se encuentre trabajando.

Este método consiste en el uso de algún método de balanceo de tipo re-muestreo sobre la clase minoritaría, como SMOTE o ADASYN, para generar un conjunto de muestras sintéticas. Posteriormente, se utiliza el método *Propensity Score Adjusted (PSA)* para generar unos *pseudo pesos* asociados a cada instancia del conjunto de muestras sintéticas para poder realizar una estimación híbrida del modelo de regresión logística que combina los datos del conjunto original y el conjunto de muestras sintéticas.

En el siguiente documento se evaluaran tres variaciones del método SMOTE para tratar el problema de balanceo de datos, los cuales son:

1. SMOTE - ENC
2. ADASYN modificado con SMOTE-ENC
3. SMOTE-ENC ajustado con pesos de muestreo

Posteriormente, se aplicará el método de PSA para generar los pseudopesos para cada conjunto de muestras sintéticas generados, finlamente se obtendran tres estimaciones del modelo de regresión logística donde se evaluara el desempeño obtenido a través de cada método de balanceo, de esta manera se podrá determinar qué variación de SMOTE junto al método de PSA se ajusta mejor al problema de balanceo de datos en la GEIH.

Balanceo

El desbalance en los datos se refiere al caso en el que el conjunto de datos no tiene un representación equitativa de instancias que pertenecen a cada clase de la variable dependiente, en este caso, se presenta un desbalance de dos clases, en donde hay una gran diferencia en el número total de instancias que pertenecen a una clase respecto a la otra, comunmente se les denomina clase mayoritaría y minoritaria.

Actualmente el total de instancias del conjunto de datos de la GEIH que pertenecen a cada clase de la variable dependiente es la siguiente:

Actividad Ocupada Ultima Semana	Frecuencia
1	5647
0	1255

Donde el valor $'0'$ indica que el joven está buscando trabajo mientras que el valor $'1'$ indica que ya se encuentra trabajando. Con esto en cuenta, se puede observar que hay un desbalance en los datos donde la clase minoritaría se refiere a la clase de los jóvenes que están buscando trabajo.

El desbalance de los datos es un tema desafiante en el machine learning. De hecho, según Mukherjee et.al (2021), "los algoritmos de machine learning tienden a predecir cualquier instancia como un elemento de la clase mayoritaria, haciendo que el modelo resulte ineficiente para identificar las instancias de la clase minoritaria, esto es algo crítico, especialmente, cuando hay un gran interes en clasificar de manera correcta esta clase".

Se han desarrollado varías formas para tratar este problema, dentro de las principales formas están los métodos de re-muestreo, esto puede ser a través del sobre muestreo de la clase minoritaría o del sub muestreo de la clase mayoritaría. Entre estos métodos de re-muestreo, uno de los métodos más utilizados es el método SMOTE (Chawla et al. 2002).

SMOTE es un algoritmo en el cual la clase minoritaría recibe un sobre muestreo a través de la creación de muestras "*sintéticas*" que se ubican en los segmentos que unen a cada instancia de la clase minoritatía con sus k vecinos más cercanos en cada variable o característica.

SMOTE ha ganado una gran popularidad entre los métodos que existen para tratar el problema de balanceo y de hecho, se ha establecido como uno de los métodos más utilizados para tratar este problema. Además, desde su desarrollo han salido multiples variantes cómo los con Borderline-SMOTE, ADASYN, SMOTE ENN, entre otros, que buscan mejorar su rendimiento en diferentes escenarios.

Ahora, un problema que tienen estos métodos, es que fueron desarrollados bajo la consideración que todas las variables son continuas, en el caso que se trabaje sobre un conjunto de datos que contiene variables nominales, como es el caso de la GEIH, tanto SMOTE como sus variantes no son directamente aplicables. Si bien existe una variante en donde se codifican las variables nominales a través de la técnica One Hot Encoding, esta variante no es la mejor solución ya que aumenta considerablemente el costo computacional del algoritmo y además, es posible que el algoritmo no aprenda sobre las posibles relaciones entre los valores nominales y las clases.

Por este motivo, se han desarrollado variantes de SMOTE que permiten manejar variables nominales y continuas. Una de estas variantes es SMOTE-ENC(SMOTE Encoded Nominal and Continuous) (Mukherjee and Khushi 2021), en donde las variables nominales son codificadas como variables numéricas y en donde un valor más alto representa una asociación más fuerte con la clase minoritaría.

SMOTE-ENC es una alternativa que nos permite tratar el problema del desbalance en conjuntos con variables numéricas y nominales, sin embargo, cabe resaltar que este método proviene de SMOTE y por lo tanto, puede heredar algunas de las limitaciones de este método.

En ese sentido, ADASYN (Adaptive Synthetic) (He et al. 2008) surge como una de las variantes de SMOTE más robusta. Este método se basa en la idea de generar muestras sintéticas de la clase minoritaría de forma adaptativa, es decir, busca generar un mayor número de muestras sintéticas de aquellas intancias con una menor densidad.

La mayor diferencia entre SMOTE y ADASYN radica en que SMOTE genera la misma cantidad de registros sintéticos para cada muestra de la clase minoritaria, mientras que ADASYN provee un peso a cada registro de la clase minoritaria para determinar el número de muestras sintéticas que deben ser generadas por cada registro.

Teniendo en cuenta que ciertas variantes de SMOTE, incluyendo ADASYN, funcionan bajo la consideración de que todas las variables son numéricas, para tratar este problema, se propone realizar una variante del algoritmo ADASYN, donde se planea utilizar la métrica dispuesta en el método SMOTE-ENC para encontrar los k vecinos más cercanos de cada instancia de la clase minoritaria. La utilización de esta métrica permite combinar la robustez de ADASYN con la capacidad que tiene SMOTE-ENC para tratar variables numéricas y nominales.

Esta variante de ADASYN con SMOTE-ENC será el segundo método utilizado dentro del documento para tratar el problema de desbalance de datos en la GEIH, se generará el conjunto de muestras sintéticas sobre

el mismo conjunto de entrenamiento utilizado en SMOTE-ENC.

Además de los métodos SMOTE-ENC y ADASYN, este documento propone una tercera alternativa para tratar el problema del desbalance de datos, esta alternativa está basado en el algoritmo WSMOTE (Prusty, Jayanthi, and Velusamy 2017).

WSMOTE es un método de sobre muestreo que asigna unos pesos a cada instancia y que determinan el número de muestras sintéticas que se van a generar a través de SMOTE para cada instancia de la clase minoritaria.

La variante propuesta consiste en adaptar la idea del uso de unos pesos para determinar el número de muestras sintéticas, sin embargo, en el caso propuesto, se elimina el calculo de los pesos explícitos, en su lugar, se van a utilizar los pesos de muestreo asociados al conjunto de datos para determinar el número de muestras sintéticas que se van a generar de cada instancia. Se propone además, para poder tener un número adecuado de muestras sintéticas, normalizar los pesos de muestreo.

Esta propuesta se basa en la idea de que los pesos de muestreo representan el porcentaje de la población que representa cada registro dentro del conjunto de datos. Utilizar estos pesos normalizados, permite generar un conjunto de muestras sintéticas que procura preservar las distribuciones asociadas a la población original.

Otra consideración es que WSMOTE es una variante que determina la cantidad de muestras sintéticas que se van a generar de cada instancia, pero, genera estas muestras a través de SMOTE, por lo cual, trabaja bajo la consideración que todas las variables son numéricas, en ese sentido, la propuesta de SMOTE con pesos de muestreo también utilizará la metrica de SMOTE-ENC para encontrar los k vecinos más cercanos.

Ahora, un detalle muy importante a tener en cuenta es que el conjunto de datos de la GEIH, es que este es un conjunto de datos basado en un muestreo complejo, por lo que cada instancia tiene asociado un peso de muestreo, que resulta fundamental al momento de realizar cualquier tipo de análisis. Este valor permite ajustar los resultados del conjunto de datos para obtener estimaciones más precisas sobre la población de estudio, y de hecho, ignorar esta variable puede llevar a realizar estimaciones imprecisas.

La solución al problema de desbalance de datos no es ajena a la consideración de los pesos de muestreo, desde los métodos planteados de re-muestreo, el resultado de estos ejercicios es el de un conjunto de entrenamiento que tiene asociados unos pesos de muestreo, y un conjunto de muestras sintéticas, que por su naturaleza se ajustan a la población de estudio, pero que no están determinados por un proceso de muestreo y por lo tanto no tienen un peso de muestreo asociado.

Para resolver este problema, se propone utilizar la idea propuesta por Elliott (2009) donde se busca generar una estimación híbrida a partir de la combinación de una muestra probabilística y una muestra no probabilística.

Se trata de un método que permite construir unos pseudo pesos para la muestra no probabilística a través del Propensity Score (Ebrahim Valojerdi and Janani 2018), una técnica que intenta estimar la probabilidad que un sujeto pertenezca a un grupo de tratamiento en función de sus covariables, y se expresa cómo:

$$e_i = P(T_i = 1|X_i),$$

asumiendo que T es pertenecer al grupo de tratamiento y X es el conjunto de covariables.

En el caso de una muestra probabilística y no probabilística, si se define a S como un indicador para saber si un elemento de la población pertenece a la muestra probabilística y S^* como el indicador que permite conocer si un elemento de la población pertenece a la muestra no probabilística, a partir de un conjunto de covariables W. Entonces el Propensity Score para conocer si un elemento de la población pertenece a la muestra no probabilística es:

$$P(S^* = 1|W)$$

Y a través de las estimaciones de esta probabilidad sobre la población, es posible obtener unos pseudopesos a través del valor $1/\hat{P}(S^* = 1|W)$.

Más aún, si definimos a Z como un indicador dentro de ambas muestras que identifica si un elemento pertenece a la muestra no probabilística $Z = 1$. Entonces podemos tener la probabilidad de $P(Z = 1|W)$, y para una muestra lo suficientemente grande vamos a tener:

$$\frac{P(W|S* = 1)}{P(W|S = 1)} \propto \frac{P(Z = 1|W)}{P(Z = 0|W)}$$

Por lo que podemos aproximar el Propensity Score a través de la estimación de la probabilidad del indicador Z, que puede ser estimado a través de un modelo de regresión logística.

Así pues, en esta regresión se ajustan los pesos iniciales para las instancias no probabilísticas con un valor de 1, mientras que las instancias de la muestra probabilística usaran sus pesos ajustados. Así, la inversa de la probabilidad resultante para las instancias no probabilísticas serán el *pseudopeso* asocidado a la muestra no probabilística.

Con esto en cuenta, si se toma al conjunto de entrenamiento como la muestra probabilística y al conjunto generado de muestras sintéticas como la muestra no probabilísitica, a través del Propensity Score, es posible estimar unos pseudo pesos de muestreo para el conjunto sintético que permita realizar un balanceo del conjunto de datos bajo la consideración de los pesos de muestreo.

Es muy importante tener en cuenta que el propensity score es consistente bajo la suposición que las muestras asociadas a cada grupo de análisis deben estar bajo un soporte común, es decir, ambas muestras deben cubrir la misma porción de la población.

Una de las soluciones que existen para asegurar que haya un soporte común entre las muestras en el *matching*, donde se selecciona de forma aleatoria un elemento del conjunto, así pues, este método garantiza que ambas muestras compartan una distribución similar dentro de las covariables del modelo, lo que permite que se cumpla con el supuesto del soporte común.

Con esto en cuenta, a través de la librería de R MatchIt, se realiza un proceso de matching a tráves del método *nearest neighbors*, finalmente se procede a ajustar un modelo de regresión logística utilizando las muestras resultantes del proceso de matching dentro de todo el conjunto de covariables del modelo.

Los pesos resultantes así como los pesos de la muestra probabilística deben de ser ajustados o calibrados según los totales de la población, según Elliott (2009), es posible realizar esto a partir de la siguiente formula:

- Para la muestra no probabilística: $\hat{w_i} = C_{s^*} \times \tilde{w_i}$ donde: $C_{S^*} = \frac{n_{s^*}}{n_{s^*}+n_s} \cdot \frac{\sum_{i \in S} w_i}{\sum_{j \in S^*} \tilde{w_j}}$
- Para la muestra probabilística $\hat{w_i} = C_s \times w_i$ donde: $C_S = \frac{n_s}{n_{s^*}+n_s}$.

Estimación del modelo de regresión logística

Para evaluar el rendimiento que tiene los métodos propuestos de re-muestreo y generación de pseudo pesos, se va a estimar un modelo de regresión logística que permita determinar si un joven se encuentra trabajando $Y = 1$ o está buscando trabajo $Y = 0$ en función de sus covariables.

El rendimiento de los modelos de clasificación, son usualmente evaluados a través de una *matriz de confusión*, una tabla en donde las filas representan los valores estimados y las filas representan los valores reales de la siguiente manera:

Table 2: Matriz de Confusión

	Valor Negativo	Valor Posítivo
Predicción Negativa	Verdadero Negativo (VN)	Falso Negativo (FN)

	Valor Negativo	Valor Posítivo
Predicción Positiva	Falso Positivo (FP)	Verdadero Positivo (VP)

A través de la matriz de confusión se pueden obtener:

- **Verdadero Negativo (VN):** El número de instancias negativas que fueron clasificadas de forma correcta.
- **Falso Negativo (FN):** El número de intancias positivas que fueron clasificadas de forma errónea.
- **Falso Positivo (FP):** El número de instancias negativas que fueron clasificadas de forma errónea
- **Verdadero Positivo (VP):** Número de intancias positivas que fueron clasificadas de forma correcta.

A partir de la matriz de confusión, se pueden cálcular ciertas métricas de rendimiento de clasificación como los son:

Accuracy: que se entiende como la proporción del total de instancias que fueron clasificadas de forma correcta y está definida como Accuracy= $(VP + VN)/(VP - FP - FN + VN)$.

Precision: la precisión intenta determinar la proporción de estimaciones positivas que efectivamente eran positivas, está definida como: Precision= $VP/(VP + FP)$.

Recall o Sensibility: se entiende como la proporcion de intancias positivas que fueron estimadas como positivas, se deine como Recall= $VP/(VP + FN)$.

Specificity: conocida como tasa de verdaderos negativos, indica la proporción de instancias negativas que fueron estimadas como negativas, se define como: Specificity= $VN/(VN + FP)$.

Balanced Accuracy: esta es una métrica que tiene en cuenta la precisión sobre ambas clases, se define como la media aritmética entre la sensibilidad y especificidad, Balanced Accuracy = (Sensitivity + Specificity)/2.

F1-Score: se puede explicar como la media armónica entre la Precisión y el Recall, por lo que permite evaluar a través de una única métrica el resultado de las métricas involucradas, esta se define como $F1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$.

Con estas métricas en cuenta, se va a estimar un modelo de regresión logística sobre el conjunto de datos original sin ningún tipo de balanceo y dos modelos de regresión logística sobre cada método de balanceo propuesto, uno de estos modelos sobre el conjunto con los pseudo pesos ajustados y otro tomando el peso de muestreo para las muestras sintéticas con el valor de 1.

Resultados

Una vez se han ajustado los modelos de regresión logística y cálculado las métricas dispuestas sobre un conjunto de prueba, los resultados obtenidos son los siguientes:

Table 3: Resultados Métodos, (**) índica el valor más alto por métrica.

Método	Accuracy	Precision	Recall	Specificity	Balanced Accuracy	F1 - Score
Conjunto Original	0.810**	0.825	0.974**	0.090	0.532	0.893**
SMOTE-ENC sin PSA	0.810**	0.825	0.973	0.094	0.534	0.893

Método	Accuracy	Precision	Recall	Specificity	Balanced Accuracy	F1 - Score
SMOTE-ENC con PSA	0.726	0.845	0.816	0.332	0.574	0.829
ADASYN-ENC sin PSA	0.810**	0.825	0.973	0.093	0.534	0.893
ADASYN-ENC con PSA	0.740	0.837	0.847	0.273	0.560	0.843
WSMOTE-ENC sin PSA	0.810**	0.827	0.973	0.095	0.534	0.893
WSMOTE-ENC con PSA	0.754	0.850**	0.848	0.340**	0.594**	0.849

El análisis revela que aquellos los modelos ajustados sobre los conjuntos de datos que no tienen un tratamiento de balanceo de datos o un ajuste de los pesos de muestreo (Original, y modelos sin ajuste por PSA) tienden a clasificar todas las intancias sobre la clase mayoritaria haciendo a los modelos insuficientes para predecir instancias de la clase minoritaría, esto se puede evidenciar en métricas como el Recall donde tienen valrores muy cercanos a 1 y la métrica Specificity con valores cercanos a 0, por lo que hay los modelos no son capaces de aprender sobre la clase minoritaría.

Los modelos ajustados sobre los conjuntos de datos con un tratamiento del desbalance junto al ajuste de los pseudo pesos, sacrifican un poco la capacidad que tienen para predecir instancias de la clase mayoritaría, pero, aumentan su capacidad para poder predecir intancias de la clase minoritaría, además, a través de la métrica Balanced Accuracy se puede evidenciar que a nivel general, estos modelos ganan más en su capacidad de aprender de la clase minoritaría que lo que pierden de capacidad frente a la clase minoritaría en comparación con los modelos de los conjuntos sin ajuste.

Dentro de los modelos sobre los conjuntos con ajuste, el método que tienen los mejores resultados, es el método propuesto de WSMOTE con pesos de muestreo bajo la métrica SMOTE-ENC, el modelo de este método de balance es aquel que tiene una mayor proporción de instancia de la clase minoritaria correctamente clasificadas, es el modelo con una mejor Balanced Accuracy, es decir, una mejor proporción de instancias de cada clase clasificada correctamente y de hecho, es el modelo cuya proporción de instancias clasificadas dentro de la clase mayoritaría pertenecian a la clase mayoritaria.

Conclusiones

El problema de desbalance de datos es un desafio bastante comun en el machine learning, si no se considera este problema, los modelos ajustados pueden ser insuficientes para aprender correctamente sobre las características de cada clase y, de hecho, es posible que al momento de evaluar estos modelos, tengan buenos resultados sobre las métricas relacionadas con la clase mayoritaría, pero esto ocurre debido a que en realidad están clasificando a todas las instancias dentro de la clase mayoritaría.

Existen diversos métodos que tratan el problema del desbalance, en este documento se tratan algunos métodos basados en el re-muestreo sobre la clase minoritaría, sin embargo, en conjuntos basados en muestreo, es necesario realizar un análisis adicional que ayude a establecer qué se debe realizar en relación a los pesos de muestreo asociados a este conjunto, en este documento se propone estimar y ajustar unos pseudo pesos sobre el conjunto sintético a través del propensity score, y según los resultados, esta perspectiva permite mejorar considerablemente la capacidad que tienen los modelos sobre un conjunto de datos basado en muestro para identificar las intancias de cada clase en la variable dependiente.

Referencias

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June): 321–57. https://doi.

org/10.1613/jair.953.

Dever, J. 2018. "Combining Probability and Nonprobability Samples to Form Efficient Hybrid Estimates: An Evaluation of the Common Support Assumption." In *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, 1–15.

Ebrahim Valojerdi, Ameneh, and Leila Janani. 2018. "A Brief Guide to Propensity Score Analysis." *Medical Journal of the Islamic Republic of Iran*, September, 717–20. https://doi.org/10.14196/mjiri.32.122.

Elliott, Michael R. 2009. "Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights." *Survey Practice* 2 (6): 1–7. https://doi.org/10.29115/sp-2009-0025.

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE. https://doi.org/10.1109/ijcnn.2008.4633969.

Mukherjee, Mimi, and Matloob Khushi. 2021. "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features." *Applied System Innovation* 4 (1): 18. https://doi.org/10.3390/asi4010018.

Prusty, Manas Ranjan, T. Jayanthi, and K. Velusamy. 2017. "Weighted-SMOTE: A Modification to SMOTE for Event Classification in Sodium Cooled Fast Reactors." *Progress in Nuclear Energy* 100 (September): 355–64. https://doi.org/10.1016/j.pnucene.2017.07.015.

Anexo

Suponga

SMOTE

SMOTE-ENC

ADASYN

Propuesta WSMOTE with Survey Weights

Código realizado en R