Comparación entre regresión logística y *random forest* para determinación de factores de violencia de pareja en el Perú

Ashley Mercedes Guerrero-Muguerza

La violencia de pareja es una problemática social que ha sido estudiada por diferentes investigadores para los factores que influyen en la ocurrencia de la misma, considerando diferentes entornos, tiempos y locaciones. El 68,2 % de mujeres han sido víctimas de violencia, y el 31,7 % fueron víctimas de agresión física en el Perú. La presente investigación propone nueve modelos basados en logística y *random forest* con las de *chi-square*, entropía y Gini, y tres sub escenarios de cinco, diez y veinte variables que utilizaron el *dataset* de denuncias registradas en el año 2016 del Ministerio de la Mujer. Se obtuvo el mejor resultado de cada subescenario, pero finalmente el mejor modelo fue el de veinte variables utilizando el *feature selection random forest (entropy)* y el modelo *random forest* (Gini).

Comparison Between Logistic Regression and Random Forest for Determining Factors of Intimate Partner Violence in Peru

Intimate partner violence is a social problem that has been studied by different researchers to determine the factors that influence its occurrence, considering different environments, moments and locations. Sixty-eight point two percent (68.2%) of women have been victims of violence and 31.7% have been victims of physical aggression in Peru. The present research proposes nine models based on logistic regression and random forest with variants such as chi-square, Entropy and Gini, and three subscenarios out of five, ten and twenty variables that used the dataset of complaints registered in 2016 at the Ministry of Women. The best result of each subscenario was obtained, but finally the best model was that of twenty variables which used the random forest "feature selection" (Entropy) and the random forest model (Gini).

Comparación entre regresión logística y *random forest* para determinación de factores de violencia de pareja en el Perú

Ashly Mercedes Guerrero-Muguerza
20150626@aloe.ulima.edu.pe

RESUMEN: La violencia de pareja es una problemática social que ha sido estudiada por diferentes investigadores para determinar los factores que influyen en la ocurrencia de la misma, considerando diferentes entornos, tiempos y locaciones. El 68,2 % de mujeres han sido víctimas de violencia, así como el 31,7 % lo fueron de agresión física en el Perú. La presente investigación propone nueve modelos basados en regresión logística y *random forest* con las variantes de *chi-square*, entropía y Gini, y tres subescenarios de cinco, diez y veinte variables que utilizaron el *dataset* de denuncias registradas en el año 2016 del Ministerio de la Mujer. Se obtuvo el mejor resultado de cada subescenario, pero finalmente el mejor modelo fue el de veinte variables utilizando el *feature selection random forest* (entropy) y el modelo *random forest* (Gini).

Introducción

Actualmente entre los diversos problemas sociales sobresale la violencia de pareja, el cual es un tipo de violencia de género, como el más grave. Este crimen es categorizado según la forma de violencia o acción cometida.

- Violencia física es la realización de lesiones, fisuras, contusiones, entre otros daños, a la pareja. El 31,7 % de mujeres han sufrido este tipo de agresión en el Perú.
- Violencia psicológica consiste en dominar y confinar a la víctima, en contra de su voluntad, y causarle daños psíquicos. El 64,2 % de mujeres peruanas han vivido situaciones que involucran este tipo de agresión.
- Violencia sexual ocurre cuando se realiza cualquier tipo de coacción con fines sexuales en contra de una persona sin su consentimiento. El 6,6 % de las mujeres en el Perú han reportado ser víctimas de esta agresión. (INEI, 2017)
- Violencia económica o patrimonial consiste en realizar un deterioro económico o financiero a la pareja (economía y patrimonio). El 68,2 % de mujeres peruanas han sufrido este tipo de agresión en el Perú. (INEI, 2017)

Materiales y métodos

Preprocesamiento de datos

La limpieza de datos consiste en utilizar diversas técnicas para estandarizar la data, en otras palabras, eliminar variables identificadas con un determinado porcentaje de nulos o variables que poseen datos nulos, pero son factibles para el modelo por lo que se deben utilizar técnicas para poder completarlos. En este caso se utilizó la media para variables como la edad, los años, y la moda para variables categóricas.

Identificación de factores influyentes

Se utilizó *feature selection* para obtener los factores que están más relacionados a la variable predictora la cuál es la violencia física. Se realizaron tres técnicas, la primera fue *chi-square* la cual pertenece al grupo de los métodos de filtro (*filther method*), la segunda y tercera fueron el *random forest* con el criterio de división de nodos de Gini y entropía, respectivamente. Por cada técnica se realizaron diferentes subescenarios de cinco, de diez y de veinte variables; de los que se obtuvieron nueve escenarios diferentes.

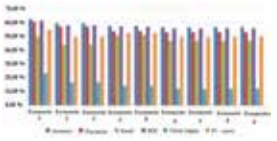
Datos

Los *datasets* utilizados se obtuvieron del Ministerio de la Mujer y Poblaciones Vulnerables. Este registro pertenece a las denuncias realizadas a nivel nacional, las cuales presentan información de las víctimas o personas que conocían a víctimas de violencia durante el año 2016. De las denuncias presentadas se seleccionaron, únicamente, las relacionadas con la violencia física de pareja.

Resultados

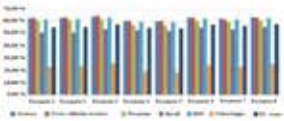
En la tabla se pueden observar los nueve escenarios realizados:

Comparación: Grupo de cinco variables



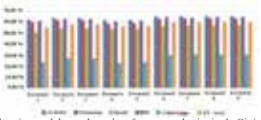
- El mejor modelo es la regresión logística con *chi-square* (escenario 1).
- ([2647, 974], [1528, 1499]).
- Cohen´s kappa: 0,2297.
- Tomando como referencia la matriz de confusión, el valor de Cohen´s kappa es el adecuado, debido al valor de los clasificados como violencia física.
- La regresión logística tiene una *performance* buena al utilizar *random forest*.
- Los escenarios 6 y 7 son los que poseen la más baja *performance*.

Comparación: Grupo de diez variables



- El mejor modelo es el de la regresión logística con el *feature importance* utilizando el criterio Gini.
- ([2624, 997], [1418, 1609]).
- Cohen´s Kappa: 0,2597.
- Tomando como referencia la matriz de confusión, el valor de Cohen´s kappa es el adecuado, **como resultado del** valor de los clasificados como violencia física.
- El segundo mejor es el escenario 8, seguido de cerca por el escenario 6.
- Utilizar *chi-square* con *random forest* no ayuda en la *performance* del modelo.

Comparación: Grupo de veinte variables



- El mejor modelo es el *random forest* con el criterio de Gini para la división de nodos y la selección de variables importantes con el criterio de *entropy* (escenario 9).
- ([2627, 994], [1264. 1763]).
- Cohen´s kappa: 0,3047
- Tomando como referencia la matriz de confusión, el valor de Cohen´s kappa es el adecuado, debido al valor de los clasificados como violencia física.
- El segundo mejor es el escenario 7, en el cual sus valores son parecidos al escenario seleccionado.
- El uso de *random forest* y la regresión logística con *chi-square* se encuentran entre los peores resultados.

- El modelo seleccionado fue el de las veinte variables, debido a que la mayoría de los factores obtenidos para dichos subescenarios se encontraron en la literatura revisada.
- Las variables más importantes del modelo son la edad de la víctima y del agresor, el tiempo de duración de la agresión, el nivel educativo de ambas partes y el número total de hijos. En comparación a los últimos cuatro factores listados, los cuales no son factores que tienen una gran importancia en el modelo, según los resultados obtenidos por el *random forest*.

Conclusiones

El uso de la regresión logística, ya sea con *chi-square* o *random forest*, para la selección de características, es recomendable dados sus buenos resultados. Sin embargo, se debe utilizar cuando los factores independientes no sean de gran cantidad, pues se observó que durante la investigación, cuando la cantidad de factores era mayor, el modelo que ofrecía mejores resultados era el *random forest*.

Cabe recalcar que, para seleccionar el mejor modelo, no se debe utilizar el *accuracy* como única métrica para ver la *performance* del modelo, debido a que a) cuando los datos están desbalanceados esta métrica no aporta mucho en el resultado y b) utilizar diversas métricas conduce a una mejor comprensión del modelo utilizado.

Referencias

- Abramsky, T., Watts, C. H., García-Moreno, C., Devries, K., Kiss, L., Ellsberg, M., … y Heise, L. (2011). What factors are associated with recent intimate partner violence? Findings from the WHO multi-country study on women's health and domestic violence. *BMC Public Health*, *11*(1), 109.
- Alves, L. G., Ribeiro, H. V., y Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, *505*, 435-443.
- Belgiu, M., y Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31.
- Berk, R. A., Sorenson, S. B., y Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, *13*(1), 94-115.
- Hsieh, T. C., Wang, Y.-H., Hsieh, Y.-S., Ke, J.-T., Liu, C.-K., y Chen, S.-C. (2018). Measuring the unmeasurable - a study of domestic violence risk prediction and management. *Journal of Technology in Human Services*, *36*(1), 56-68. doi:10.1080/15228835.2017.1417953
- Instituto Nacional de Estadística e Informática (INEI) (2017). *Perú: Indicadores de violencia familiar y sexual, 2000-2017*.
- Izmirli , G., Sonmez, Y., y Sezik, M. (2014). Prediction of domestic violence against married women in southwestern Turkey. *International Journal of Gynecology and Obstetrics*, *127*(3), 288-292.
- Kraanen, F., Vedel, E., Scholing, A., y Emmelkamp, P. (2014). Prediction of intimate partner violence by type of substance use disorder. *Journal of Substance Abuse Treatment*, *46*(4), 532-539.
- Saile, R., Neuner, F., Ertl, V., y Catani, C. (2013). Prevalence and predictors of partner violence against women in the after math of war: A survey among couples in Northern Uganda. *Social Science & Medicine*, *86*,17-25.

Agradecimientos

Agradezco a mis profesores, quienes en todo momento me alentaron y apoyaron en el desarrollo del presente trabajo; especialmente a los ingenieros Juan Gutiérrez, Vilma Romero, Rosario Guzmán y Pablo Rojas, y a la magíster Rosa Millones.