Factores de percepción de terremotos en Argentina

Trabajo final de la asignatura "Taller de tesis"

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

Víctor A. Bettachini

16 de junio de 2024

Resumen

Niente.

Índice general

Capítulo 1

Introducción

1.1. Contexto y motivación científica

Un rápido desprendimiento entre dos fascetas enfrentadas de sendas placas tecnónicas que trabanban mutuamente su desplazamiento relativo produce una rápida liberación de energía que se denomina terremoto. Esto sucede a cierta profundidad en la corteza terrestre en el punto denominado hipocentro a partir del cual parte de esta energía se encauza como ondas elásticas. El estudio de estas ondas es el área llamada sismología y de ahí el termino sismo para un evento particular detectado, pero que debiera aclararse si se produjo por un terremoto u otra fuente de ondas [1, sección 4.1.1]. Sea que las ondas sean de compresión longitudinal de la corteza, las tipo P, o las del tipo S transversales y más lentas, ambas arriban con mayor intensidad al punto de la superficial terrestre que se encuentra directamente sobre el hipocentro, que se denomina epicentro [1, sección 4.1.2] como ilustra la figura 1.1. Cuanto más proxima es una locación en la superficie a un epicentro la amplitud de las ondas sísmicas es mayor. Tanto esta amplitud como el período de oscilación son mucho mayores que el de otros desplazamientos de la corteza respecto a otros desplazamientos como los de las mareas solares y lunares de la corteza [1, sección 4.1.4]. Como resultado estructuras artificiales pueden agitarse poniendo en riesgo su estabilidad estructural y haciendo caer elementos que no estaban fijados a esta o perdieron tal adhesión a causa de la agitación misma. Como consecuencia los sismos más fuertes pueden generar graves daños, poniendo en riesgo la integridad física y la seguridad de las personas al generar daños en las viviendas y edificios, derrumbes de puentes, rompimiento de vidrios, entre otros [2].

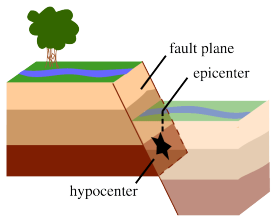

Figura 1.1: Epicentro e hipocentro de un sismo.

Pero de hecho la mayor parte de los sismos son de muy pequeña magnitud y no generan daños materiales. Sin embargo el que sean o no detectados por la población es un factor relevante en su percepción de confianza vis-à-vís de los organismos de monitoreo y prevención de riesgos. Sí o cuando informar a

la población de la ocurrencia de un sismo es una decisión de política pública que debiera apuntar a no alertar innecesariamente sobre sismos menores imperceptibles [3]. El caso inverso es también problemático con casos en que ante una carencia de una comunicación oficial de la poca importancia de un evento sismico llevó a la autoevacuación por parte de la población que lo percibió [4].

1.2. Objetivos del trabajo / Pregunta

Contar con una estimación rápida a partir de los datos sismicos registrados por instrumental de si un dado evento será percibido por la población o no permitiría a las autoridades tomar decisiones informadas sobre la comunicación a la población. Este trabajo busca determinar si con métodos de ciencias de datos puede predecirse con un grado de certeza elevado si un sismo será percibido por la población o no a partir de las variables de los sismos que se registran en los datos del Instituto Nacional de Prevención Sísmica (INPRES) que es el organismo público de Argentina que realiza estudios e investigaciones básicas y aplicadas de sismología [5].

1.3. Estructura del documento

Se estructuró en los siguientes capitulos con temáticas difenciadas

Introducción : se presentará el contexto y la motivación científica, los objetivos del trabajo y la estructura del documento.

Marco teórico : se revisarán los trabajos previos y relevantes, se presentarán los conceptos y técnicas de ciencia de datos utilizados en el trabajo.

Metodología : se describirán los datos utilizados, el preprocesamiento y limpieza de los mismos, el análisis exploratorio de los datos, las técnicas de análisis y modelado utilizadas, la selección de características, las métricas de evaluación de los modelos y los métodos estadísticos utilizados.

Resultados y discusión : se presentarán y analizarán los resultados obtenidos, se discutirán los resultados y su relevancia, se identificarán las limitaciones y posibles mejoras.

Conclusión : se resumirán los hallazgos principales, se presentarán las conclusiones generales y su relación con los objetivos del trabajo, se discutirán las aplicaciones y relevancia de los resultados.

Capítulo 2

Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo

Siendo el objetivo de cualquier modelo utilizado una clasificación entre dos clases, la percepción de un sismo o no, se busca determinar que tan fuerte es el vínculo de cada una de las variables de los sismos de las que se dispone datos y la percepción de los mismos por parte de la población. Para esto dos técnicas se planea utilizar dos distintas técnicas de la ciencia de datos: la regresión logística y la clasificación de árboles de decisión.

Un modelo de regresión logística permite una clasificación binaria de los datos en función de una variable dependiente y un conjunto de variables independientes al tiempo de dar un peso a cada una de estas últimas lo que permitirá transmitir cuales son más relevantes en la clasificación. Determinar la atribución de la percepción de un sismo a las variables de los mismos se considera de interés para los objetivo de este trabajo. Ajustando el grado de regularización tipo Lasso (L1) se buscará reducir el número de variables independientes en el modelo verificando en que grado su omisión afecta a la predicción.

Por otra parte se buscará utilizar una herramienta más reciente, las máquinas de potenciacion de gradiente, más conocidos por su nombre en inglés Gradient Boosting Machines (GBM). Las distintas implementaciones de estos algoritmos, como XGBoost, LightGBM o CatBoost son capaces de producir un único modelo con fuerte poder predictivo a partir de la síntesis de resultados de modelos de predicción débiles, típicamente árboles de decisión.

Puesto que los GBM carecen de un mecanísmo para evidenciar la importancia de las variables en la clasificación como la que evidencian los pesos de la regresión logística, se planea utilizar los valores de explicaciones aditivas de Shapley (SHapley Additive exPlanations, SHAP) para tal fin. Para el lenguaje R está disponible la biblioteca *shapr* para tal fin [6].

El disponer de una herramienta de explicación de los modelos de predección aplicable al resultado de los dos a ensayar, el de regresión lógistica y el XGBoost,

permitirá comparar la relevancia de las variables que asignará cada uno para la percepción de los sismos por parte de la población.

Capítulo 3

Metodología

3.1. Presentación y descripción de los datos utilizados

En el marco de los "Proyectos de Asistencia Estadística del Instituto de Cálculo (IC)" de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad de Buenos Aires (UBA) se publicaron conjuntos de datos en un repositorio curado con el objeto de ser aplicados a la enseñanza de la estadística y la ciencia de datos por Daniela Parada, investigadora del IC [7]. De estos conjuntos el utilizado en este trabajo es el que se publica en el apartado "Visualización" que corresponden a datos de sismos de Argentina de los últimos 10 años [8]. En este repositorio alojado por la firma GitHub, se provee un front-end html que da un contexto, hace una exploración inicial, un analísis para una provincia en particular, muestra una estimación de probabilidad y provee otra información sobre lo datos.

Los datos corresponden a detecciones por parte de estaciones de monitoreo sísmico en la República Argentina recopilados y publicados por el INPRES en su sitio web [9]. En el sitio de publicación de los datos se indica que el conjunto de datos comprende las fechas desde el 7 de enero de 2012 hasta el 18 de mayo de 2022 y fue realizado con datos *scrappeados* del buscador de sismos del INPRES por Gustavo Juantorena [8, sección 4.1].

Allí mismo se describe que el conjunto de datos reducido y curado denominado "sismos", el que se utilizó en este trabajo, es accesible a través de la importación de la biblioteca `datosIC` en lenguaje R [8, sección 5.1.1]. Este mismo conjunto reducido puede descargarse en formato de texto separado por comas (CSV) apuntando a su URL en el repositorio GitHub[1].

Las variables reportadas para cada sismo son:

- *Fecha*: en el formato aaaa-mm-dd de la norma ISO 8601 [10].
- *Hora*: una cadena de caracteres en formato hh:mm:ss lo que da una preción al segundo para los datos.

[1] `https://github.com/daniellaparada/IC-datasets-docencia/blob/main/fuente/04_visualizacion/sismos-arg.csv`

- *Latitud, Longitud*: un número con una precisión de un decimal con grados como unidad.
- *Provincia*: cadena de caractéres del nombre de la provincia donde se produjo el sismo (no donde se ubicó quién potencialmente lo percibiera) según se afirma en el sitio de publicación [8, pág. 5.1.1].
- *Magnitud*: un número con la escala Richter como unidad una función de la amplitud de las ondas sísmicas [1, sección 4.2.3]
- *Profundidad*: un número entero con kilometros como unidad que indico que tan bajo la superficie se ubicó el epicentro.
- *Percibido*: lo que en el sitio del INPRES se califica como "sismos sentidos" es una clase categórica de si hubo reportes de percepeción del fenómeno por parte de la población.

Esta última variable es la que se busca predecir en este trabajo en función de las demás.

3.2. Preprocesamiento y limpieza de los datos

Puesto que el conjunto de datos es curado por un equipo de investigación de la UBA, se asume que los mismos son confiables y que no se requiere de un proceso de limpieza de los mismos. De todas formas se realizaron las verificaciones usuales cada vez que se utilizan datos tabulares en un estudio de estadístico y/o de ciencia de datos.

Inspección de columnas Tras descargar el archivo de datos en fomato CSV y importarle a un entorno de trabajo en lenguaje R en una estructura de datos "data.table" denominado `sismos_arg` ejecutar `colnames(sismos_arg)` permitió verificar que contuviera las columnas con los nombres anunciados en el sitio que publica los datos en su sección [8, Exploración inicial]. Asimismo el tipo de datos se constató ejecutando `str(sismos_arg)`.

Valores faltantes o duplicados La presencia de valores faltantes indicados con el símbolo `NA` se descartó cuando la ejecución `sum(is.na(sismos_arg))` arrojó un cero como resultado. Por el contrario se encontró un número de 23 filas duplicadas (sobre 55817 registros) ejecutando `sismos_arg[duplicated(sismos_arg, fromLast = TRUE)]`. Se hizo una copia de la tabla sin registros duplicados ejecutando `sismos_arg[!duplicated(sismos_arg)]` en una nueva tabla con nombre más corto, `sismos`.

Datos atípicos La detección y potenciales acciones sobre datos atípicos se tratan una vez iniciado un análisis exploratorio de datos temática de la sección 3.3.

Por otra parte el preprocesamiento comprende la generación de nuevas variables a partir de las existentes que se consideren relevantes para el análisis, lo que recibe el nombre de ingeniería de características (feature engineering). Se generaron dos nuevas columnas a partir de las existentes en el conjunto de datos

original, una en función al formato de los datos en la columna "Hora" y otra en función de la escala física utilizada en la columna "Magnitud".

Linealización de la magnitud La escala de Richter de un terremoto se determina a partir del logaritmo en base 10 de la amplitud de las ondas registradas por los sismógrafos incluyendo ajustes para compensar la variación en la distancia entre los diversos sismógrafos y el epicentro del terremoto [11]. Puesto que el objetivo es predecir la percepción de los sismos, es razonable que esta guarde una relación con la amplitud de las ondas. Con tal fin se generó una columan con la manginitud linealizada exponenciado con base 10 los datos de magnitud a través del comando `sismos[, magnitud_lineal := 10^Magnitud]`.

Decimalización de la hora del sismo La columna "Hora" está codificada como una cadena de caracteres. Para poder utilizarla en un análisis de regresión se la convirtió a un número entero de segundos transcurridos desde la medianoche del día en que se produjo el sismo. Para ello se escribió una función `convert_to_seconds` centrada en la función `strptime` del paquete base de R. Se verificó que la columna generada con aplicación a través de `sismos[, Hora_decimal := sapply(Hora, convert_to_seconds)]` estuvian en el rango de 0 a 86400 segundos, es decir, un día completo.

3.3. Análisis exploratorio de datos (AED)

Un primer vistazo sobre los datos con `summary(sismos)` permitió obtener un resumen de las variables numéricas y categóricas. Saltan a la vista que hay valores extremos en la variable "Magnitud" y un fuerte desbalance en la variable "Percibido", la de clase de clasificación.

Desbalance de la clase de clasificación Sobre el total de 55794 registros únicos, un $\approx 96{,}6\,\%$ solo fueron percibidos por el instrumental y no por la población. Restan tan solo unos 1905 registros, un $\approx 3{,}4\,\%$, que si fueron percibidos por la población. Este desbalance llama al uso de técnicas de balanceo de clases en los modelos de clasificación a utilizar.

Distribución de "Magnitud" La variable "Magnitud" presenta una distribución asimétrica por el hecho de que cuanto mayor es la energía liberada, más infrecuente es el sismo.

Horario de los percibidos Se realizó un gráfico de densidad de los horarios de los sismos percibidos por la población y los no percibidos.

3.4. Descripción de las técnicas de análisis y, si corresponde, de modelado

3.5. Descripción de la selección de características (si corresponde)

Evidentemente la proximidad de un epicentro a una población es un factor relevante en la percepción de un sismo. Lamentablemente en el conjunto de datos curados no se dispone de la ubicación de la población que informó haberles percibido lo que imposibilita incorporar la distancia al epicentro como una variable en el modelo. Esto no es un limitante en el registro de datos pues el INPRES tiene una metodología operativa para registrar la ubicación de los usuarios que reportan haber percibido un sismo a través de su página web [12].

3.6. Descripción de las métricas de evaluación de los modelos (si corresponde)

3.7. Descripción de los métodos estadísticos utilizados (si corresponde)

Capítulo 4

Resultados y discusión

4.1. Presentación y análisis de resultados obtenidos

4.2. Discusión de los resultados y su relevancia

4.3. Limitaciones y posibles mejoras

Figura 4.1: Figura 1.

Capítulo 5

Conclusión

5.1. Resumen de los hallazgos principales

5.2. Conclusiones generales y su relación con los objetivos del trabajo

5.3. Aplicaciones y relevancia de los resultados

Capítulo 6

Bibliografía

[1] C. M. R. Fowler. *The Solid Earth: An Introduction to Global Geophysics*. 2.ª ed. Cambridge University Press, 20 de dic. de 2004. ISBN: 978-0-521-89307-7 978-0-521-58409-8 978-0-511-81964-3. DOI: 10.1017/CBO9780511819643. URL: https://archive.org/details/solidearthintrod0000fowl_q3v4 (visitado 16-06-2024).

[2] *¿Qué es un sismo?* Sistema Nacional para la Gestión Integral del Riesgo. 12 de nov. de 2018. URL: https://www.argentina.gob.ar/sinagir/riesgos-frecuentes/sismos (visitado 16-06-2024).

[3] Saunders, J. K., Minson, S. E., Cochran, E. S., Bunn, J., Baltay, A. S., Kilb, D. y O'Rourke, C. «A Twist of PLUM: Low-Magnitude Earthquakes and Ground-Motion-Based Early Warning». En: 2021 Southern California Earthquake Center Annual Meeting, SCEC Contribution #11360. URL: https://www.scec.org/publication/11360 (visitado 17-06-2024).

[4] Sandra Vaiciulyte, David A. Novelo-Casanova, Allen L. Husker y Ana B. Garduño-González. «Population response to earthquakes and earthquake early warnings in Mexico». En: *International Journal of Disaster Risk Reduction* 72 (1 de abr. de 2022), pág. 102854. ISSN: 2212-4209. DOI: 10.1016/j.ijdrr.2022.102854. URL: https://www.sciencedirect.com/science/article/pii/S2212420922000735 (visitado 17-06-2024).

[5] *Instituto Nacional de Prevención Sísmica*. Argentina.gob.ar. 28 de oct. de 2022. URL: https://www.argentina.gob.ar/inpres (visitado 17-06-2024).

[6] Camilla Lingjærde, Martin Jullum y Nikolai Sellereite. *shapr: Explaining individual machine learning predictions with Shapley values*. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/shapr/vignettes/understanding_shapr.html (visitado 16-06-2024).

[7] *IC-datasets-docencia*. URL: https://daniellaparada.github.io/IC-datasets-docencia/ (visitado 12-06-2024).

[8] Daniela Parada. *IC-datasets-docencia - 4 Visualización*. URL: https://daniellaparada.github.io/IC-datasets-docencia/04_visualizacion.html (visitado 12-06-2024).

[9] *Buscador de sismos*. Instituto Nacional de Prevención Sísmica. URL: http://contenidos.inpres.gob.ar/buscar_sismo (visitado 15-06-2024).

[10] *ISO 8601-1:2019(en), Date and time — Representations for information interchange — Part 1: Basic rules*. International Organization for Standardization. 2019. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:8601:-1:ed-1:v1:en (visitado 16-06-2024).

[11] Willian L. Ellsworth. «Earthquake Magnitude: THE RICHTER SCALE (ML)». En: *The San Andreas Fault System, California*. Ed. por Robert E. Wallace. Vol. Professional Paper 15151. P. United States Geological Survey (USGS), 1991, pág. 177. URL: https://web.archive.org/web/20160425121745/http://www.johnmartin.com/earthquakes/eqsafs/safs_693.htm (visitado 14-10-2008).

[12] *Acerca de tu ubicación*. Instituto Nacional de Prevención Sísmica. URL: https://www.inpres.gob.ar/desktop/conoce.html (visitado 17-06-2024).

Capítulo 7

Anexos (opcionales)

7.1. Código fuente utilizado en el análisis (link a un repositorio)

7.2. Tablas y gráficos adicionales

7.3. Otros materiales relevantes